Depression Signs Detection through Smartphone Usage Data Analysis

Nino Rafael da Silva Rocha

Mestrado integrado em Engenharia de Redes e Sistemas Informáticos

Departamento de Ciência de Computadores

2014

Orientador

Ana Vasconcelos (MSc), Fraunhofer Portugal

Coorientador

Rita P. Ribeiro (PhD), DCC - FCUP

Todas as correções determinadas pelo júri, e só essas, foram efetuadas.

O Presidente do Júri,

Porto, ______/______/_________

Acknowledgements

First of all, i would like to thank to my supervisors, Rita P. Ribeiro (PhD, DCC-FCUP) and Ana Vasconcelos (MSc, Fraunhofer Portugal) for all their kindness and dedication.

To Dr. Ana Nobre, for the time spent with guidance in the field of psychology.

To Fraunhofer Portugal for giving me the opportunity to develop this project.

Next, thanks to my dear family, specially my father, mother, and sister, for believing in my skills, and for my education. All I am is because of them.

To my dear friends, specially to André Francisco, Ana Raquel Azevedo, André Gaspar, Filipe Martins, Geno Pereira, Hugo Sousa, Inês Caeiro, José Pedro, and Tiago Melo.

A special thanks to my best friend who always believes in me, God.

Resumo

A depressão é uma das doenças mentais mais comuns entre a população geral. Os sintomas de depressão são diversos: insónias, perda de energia, tristeza e isolamento. A depressão torna-se mais comum em estágios tardios da vida, alimentada principalmente pela perda de familiares e cessação de actividade. Este trabalho propõem um estudo que utiliza informação colectada a partir de *smartphones*, como a localização, ou estatísticas de comunicação, entre outros, e conduz uma análise estatística e de machine learning para inferir conclusões sobre o estado depressivo de idosos. A análise estatística provou ser útil para determinar conclusões baseadas nos vários campos, enquanto que o machine learning deriva a suas conclusões a partir de hábitos populacionais. Um serviço web foi construído sobre esta implementação como uma ferramenta de visualização de dados. Este projeto é ainda um estudo preliminar que outros poderão utilizar como base para outros trabalhos.

Abstract

Depression is one of the most common mental disorders among the general population. Depression symptoms are diverse: sleep disorder, energy loss, sadness, or isolation. Depression is most common in late stages of life, mostly fueled by the loss of relatives and retirement. This work proposes a study that uses data collected from smartphones, such as location sensors, communication statistics, among others, and uses statistical analysis and machine learning to infer conclusions on elders' depression symptoms. The statistical analysis proved useful for determining field-wise conclusions from personal habits, while the machine learning process derives its conclusions from general population standards. A web service was constructed on top of this implementation as an useful data visualization tool. This is a preliminary study which might serve as ground basis for others to build upon.

Contents

List of Figures

Listings

Chapter1

Introduction

The starting point of this project was the Smart Companion, an Android customization developed at Fraunhofer Portugal specifically for seniors[41], that already collects specific data through specially developed algorithms enabling the users to receive medication reminders, call the emergency line, making calls,messages and among others features. In order to specify the data requirements that should be analysed we interviewed two psychologists specialized in dealing with seniors.

Depression is one of the most common mental disorders among the general population and can be manifested from childhood to old age. This is a serious mental health problem that is even considered as the leading cause of disability related to illnesses and health problems. Studies show that one in four people in the world suffers or will suffer from depression [11].

Depression symptoms include, among others: insomnia or excessive sleeping, loss of energy, sadness, or isolation. These can often be ignored or mistaken with normal age-related behaviour, but with proper monitoring, early signs of depression could be detected early making it possible to promptly provide appropriate care.

Currently, smartphones and their integrated sensors provide a set of data that, with the proper analysis, can be used to build a behavioural pattern of its user. By analysing this data we propose an application that uses statistical analysis and Machine Learning to detect depression signs.

1.1 Goals

The goal of this project is to implement a web platform which helps in identifying variances in behavioural patterns that might indicate depressive symptoms, by using a Machine Learning algorithm, statistical analysis, and visualization tools. This can aid psychologists and caregivers in detecting depressive symptoms at early stages, allowing for more efficient treatments of depression. We consider important to allow the psychologist to be the one providing the training data sample, which allows comparing methodologies used by different psychologists. We also implement history visualization which also permits comparing results. As such, this application might be useful not only for early detection of depressive signs but also to improve treatment methodologies.

1.2 Thesis Structure

In the present chapter we have described the motivation and the main goals of this thesis. In Chapter 2 is presented the state of the art about depression detection and how to achieve successful aging. In Chapter 3 we explain the concepts about knowledge discovery, Data Mining, Machine Learning and, in particular, the Data Mining process that we implement in our project. In Chapter 4 we detail our research about existing tools that can be similar to our project. Chapter 5 focuses in explaining each step we took in developing our project. In Chapter 6 we introduce the technologies we used and the underlying architecture. Finally, in Chapter 7, we show our conclusions and future work.

Chapter2

Seniors and Depression in Old-age

2.1 Age-related Changes

The development and ageing processes are strictly individual and there are several conditions, happenings and changes in life that influence the whole ageing process [20]. Throughout this text we will focus essentially on changes at the perceptual, cognitive and psychosocial levels which most commonly occur to seniors .

- **Perceptual changes:** Throughout the ageing process we can witness great perceptual changes which may be functional or structural. These changes mostly interfere on the performance of daily tasks and indirectly on seniors participation in society.

  The deterioration of the sensory system level - responsible for processing sensory information - may influence and originate some difficulties on a motor level, physical and intellectual activities, and may inclusively jeopardize and compromise seniors quality of life. These sensory changes may directly affect the senses, namely: sight, hearing, touch, palate and smell. When it comes to sight, a significant decrease in the eye's transition and adaptation capacity might happen, resulting in a decreased perception of a distant object and lack of sensibility towards color interpretation. As for hearing, the reduction in listening high frequencies results in a difficulty to distinguish phenomena, in the misinterpretation of hurried verbal speeches or challenges in understanding speech in loud environments or even in parallel conversations. Touch is the sense responsible for the skin's perception and also suffers changes throughout the ageing process, namely the loss of perceptive capacities which consequently compromises the balance and motor coordination. In addition, there is a bigger tolerance to extreme pain stimulus, diminished limb positioning and its articulation. Conscience also suffers changes to its structure, since seniors present bigger difficulties in detecting the forms and sizes of objects when in contact with the body [29].

- **Cognitive Changes:** On the cognitive level, the decrease in cognitive task's performance (namely in terms of speed of information processing and resolution of more complex/new tasks) occurs on a gradual rate, being therefore seen as a normal aspect of ageing. Associated with this there are also several changes to the mental processes, such as perceptions, memory and intelligence. As such, we may point out that there is a connection and simultaneous operability between the sensory system and the cognitive system and its decline.

  Braun and Lalonde cited in [19] associate the cognitive system's decline only to the natural ageing process, since it is independent of social and health related factors. Salthouse (1989,1999) in [19] adds that the cognitive decline occurs due to the variations of the information propagation rate. For seniors, the changes in the mental processes in this stage in life may be characterized by a profile of cognitive decline, in which one can register several changes: increase in the time required to codify and store information, reduction of learning speed, difficulty in retaining information related to long term memory and difficulty in the interpretation of tridimensional pictures [6].

- **Psychosocial Changes:** Seniors deal constantly with the big challenge of overcoming adversities, most of which are related to changes at a physical level. The absence or decrease of physical and motor capacities and dexterity are transformed in a sensation of inadaptability and frustration against the environment in which the individual is. These self-depressive sensations and feelings on the individual may often cause a social detachment, in order to avoid the perception of these changes so meaningful to one's everyday life and confrontation with reality. Adding to these psychological changes there are also changes on a social level, on close social relations. The negative events and particularly the death of important people in a close circle are some of the most frequent situations at this age, which leads almost inevitably to the realization that one "is old". There are other great challenges on a social level in the stage of old-age, such as the fear of eventual isolation and the integration in some institution, or the parting of one's descendants from home.

The role and status of an individual in society may also be a subject of many important changes for seniors - retirement and widowhood are two of the situations in which this matter is most evident. In the specific matter of retirement, seniors have the impression that one's contribution on a working level is over, resulting in more free time which needs to be occupied somehow.

In addition, elders go through countless changes in different levels, mostly as consequence of physiological, psychosocial and perceptual changes. Side by side with the decline seniors face on several levels, there is still an increase of social and familiar and even personal pressure on the senior to be able to deal with these factors, adjusting one-self to one's environment and find a possible and desired balance [39].

2.2 Depression

Depression is one of the most common mental disorders among the general population and can be manifested from childhood to old age. This is a serious mental health problem that is even considered as the leading cause of disability related to illnesses and health problems. According to data from some studies [11], one in four people in the world suffers or will suffer from Depression. In Portugal, one in five Portuguese is depressed and 1200 deaths are a direct result from this illness [23].

This mental illness is essentially characterized by depressive symptoms that can be episodic, recurrent or chronic such as low mood, reduced energy and loss of interest or enjoyment. When these symptoms are detected, they are also usually associated with inactivity, physical pain, poor concentration, reduced self-confidence, pessimism, disturbed sleep and altered appetite [7] leading a person to a substantial decrease in their ability to comply with their daily responsibilities. The causes associated with the manifestation of mental illness differ from person to person. In most cases, depressive episodes result from a variety of biological, psychosocial and family factors that increase the risk of a person developing a depressive disorder.

Biological factors that result in increased risk of depression include: the females has predisposition to the illness, especially in adolescence;people who suffer from physical pain without explanation; chronic diseases such as hypertension, history of thrombosis, asthma, diabetes; or other diseases such as AIDS and Cancer. People with substance abuse such as alcohol or drugs are also prone to developing depressive symptoms [55].

At a family level, individuals who have a family history of depression, those who live with a family member carrying a serious or chronic disease and significant loss of a close relative are among the main factors for depression.

The psychological and social factors have a significant role in the development of depression mainly on the individuals with stress-generating professions, people prone to anxiety and/or panic, loss of employment and loss of social relationships in their midst.

2.2.1 Aging and Depression

Defined by the American Psychiatric Association's in Diagnostics and Statistical Manual (DMS-IV), Depresion in the Elderly is the existence of a depressive syndrome in individuals with over 65 years of age [32].

Since the worldwide trend is that of an increasingly aged population, the importance of diagnosis and treatment of depression in this age group must be highlighted. According to the World Health Organization, in 2025 there will be 1.2 billion people over 60 years of age and, in Portugal, the National Statistics Institute predicts a considerable increase in population percentage over 65 years of age, rising from 17,6% in 2008 to 32.3% in 2060 [12].

According to Melo and Ferruzzi in [34], Zimmerman (2000) stated that the probability of seniors suffering from this disease is greater than during youth or adulthood and this increase is justified by numerous losses and limitations associated with old age having consequences like low self-esteem. There are many risk factors associated with the onset of depression at this stage of life. At the psychosocial level, one of the most significant events is the end of an occupation. This can be negatively faced by the seniors since it leads to change routines and possible feelings of uselessness (Fernandez-Balesteros and Izal,1993 in [51]). The loss of loved ones and people in their social environment and the sometimes remoteness of family leads the seniors into a new cycle of life which can cause feelings of discouragement and loneliness (Figueiredo,2007 in [51]). Biological factors also pose a high risk of increased probability of depressive disorders. According to Nunes (2008) in [51] the cognitive decline associated with memory loss is one of the main complaints of people with depression. Associated with this are also functional insufficiency and physical illness that prevent seniors from participating in their usual daily activities. These factors combined can trigger feelings of inadequacy and social demotivation that are so often linked to Depression.

According to Fernandes in [26], Marques and Col(1989) summarize the factors of depression in seniors in three major areas: environmental determinants, particularly the isolation and the lack of social interaction and job, death of the spouse, and social and occupational devaluation. The genetics area considers seniors as a group with a genetic predisposition to depression. Finally the organic area refers to the wide variety of organic illnesses that may present symptoms of depressive nature. It should also be noted the importance of concomitant diseases and their respective medications that may results in side effects such as depression.

The branch of Medicine that focuses on the study of depressive symptoms in seniors is Geriatrics, aiming at the prevention and treatment of diseases at late stages of life. The monitoring and diagnosis are always done under the supervision of a doctor, but can be assessed through various evaluative scales available for patients. The most common scales cited in the literature are the Geriatric Depression Scale (GDS) and Center for Epidemiologic Studies Depression(CES-D). The GDS is a self-assessment questionnaire consisting of 30 questions with which the geriatric population can assess depressive symptoms. This same scale is used by professionals in monitoring and comparing the depressed state of the patient [37]. The CES-D has a smaller version of 10 items, characterized by a high specificity and sensitivity in the diagnosis of major depression in the hospital context [25]. Other scales used for screenings are the Hamilton Depression Scale(HAM-D) and the Patient Health Questionnaire-9(PHQ-9). The HAM-D is a standard method used in clinical studies and is not specific to the geriatric population and includes some items related to somatic pain so it may cause some confusion with chronic diseases that may arise [37]. The PHQ-9 is a 9-item questionnaire derived from the DSM-IV for major depressive disorders and it is a instrument that aims to firstly, measure the severity of the depression and on the other hand is an auxiliary diagnostic tool for major depression [37].

2.3 Successful Aging

The concept of Successful Ageing was born in the 60's and can be defined as a set of mechanisms of adaptation to the specific conditions of old age, looking to establish a balance between the capacity of the individual and the demands of the environment [20].

To study this concept, there is a need to overcome the stereotype commonly associated with this phase of life that this is a period in which the occurrence of disease is more frequent, the psychological and physical abilities decline and that there are obvious disabilities. To sustain a positive view of old age and the aging process, several studies were developed by The McArthur Foundation, such as "Study of McArthur Foundation"(1984) which is summarized in the book "Successful Ageing"(Rowe and Kahn,1998), which got a huge impact on the scientific community specialized in this area [50]. Rowe and Kahn(1998) argue that successful ageing, is based on "several factors that allow individuals to continue to work effectively, physically and mentally in old age [20]."

However, there is no specific pattern of successful ageing as this is a complex construct. Baltes and Cartensen (1996) in [20] state that there is no theory or standard criteria to defining success in old age but there are two related processes. On the one hand, the ability to adapt to losses that occur in old age, and on the other hand, the selection of certain lifestyles in order to maintain a good physical and mental activity. According to this authors, we can say that there is not a single path to reach the success in ageing, but a variety of factors that have a huge importance.

Margoshes (1995) states that seniors should make the management of the available time in a more conscious and balanced way, transforming their lifestyle. In this conscious and balanced way are included a positive mental attitude, continuous challenges, cognitive exercises and preservation of healthy habits, thus ensuring the success of aging [20]. To Rowe and Kahn (1998), the low risk and disability related to diseases, mental and physical functioning/active engagement with life are the 3 components able to provide successful aging, concluding that successful ageing is dependent on choices and on the social behaviors that can be obtained through individual effort [20].

All theories of successful ageing see the individuals as pro-actives and able to regulate their quality of life by setting goals and work to achieve them. For seniors to maintain a successful ageing, they must adjust to age-related changes and involve themselves actively to preserve their well-being. Active ageing or successful ageing, corresponds to the adoption of appropriate strategies to deal with the inherent challenge of the ageing process. Considering that there is no single path to grow old, it is inevitable to state the absence of a similar way to all people, because the pathways of aging are all different and it is possible to reach satisfaction and success in life through different routes. Noting the complexity to outline a pattern in this personal success in aging, it is of most importance to evaluate several criteria:

The Competence reveals itself of high importance because it is possible to predict the psychological state of individuals through this criterion. Each individual adapts in a dynamic way to the biological aging and all changes in social network and may say that aging is successful as greater is the adaption to changes. Paúl (2001), affirms that it is necessary to pay attention to the bio-psychosocial and behavioral complexity of seniors, properly valuing individual responses which are most appropriate to counter hypothetical losses of competence that jeopardize the autonomy of the subject. The integrity of autonomy in elderly life is implicit for successful aging, it is extremely necessary and required to have physical exercise and a constant existence of social relationships [20].

Health Promotion is one of the aspects that have influence on successful ageing. According to Rowe and Kahn (1998), there should be prevention in adulthood emphasizing the influence of healthy lifestyle and health self-rated and how they reveal themselves in general well-being during aging [42].

Cognitive Activity is another criterion present to obtain a good aging. The adoption of compensatory measures to cope with an expected unfavorable evolution of biological variables such as sensory loss and decreased speed of information processing, emerges as an essential combat factor to the fatalist view that old age corresponds to the loss of capacity to understanding and learning, highlighting as preventive measures the physical exercise and cognitive training. Finally is it noted the importance of adopting strategy selections, compensation and optimization in relation to cognitive decline. Baltes and Cartensen (1999) in [20], describe this strategy in a distribution of available cognitive resources for the needs and goals to which the senior assigns more importance.

Psychological Well-being is one of the central aspects of aging successfully. The competence, socioeconomic status and social integration emerge as the most important factors in measuring satisfaction and well-being. On the other hand, losses in areas such as retirement, widowhood and health issues, result in a negative impact on life satisfaction and psychological well-being [30].

The Context of Residence plays an equally important role to understanding the different patterns of aging and explain why some people achieve successful aging. This satisfaction is explained by the theory of person-environment fit (Kahana and Lawton,1992) in [20], which provides a satisfactory adaptation when in the transactions between the person and the environment, the individual characteristics of a person are congruent with the demands of the environment. The notion of aging-in-place is central to understanding the relationship between the context of residence and successful aging. Therefore, it is necessary to provide opportunities for older people to get a relationship with other people and find someone that they can trust, this being the best antidote against loneliness.

The Quality of Life, is one of the criteria of the success formula at this age. The commonly indicators used to assess this criterion are focused on subjective well-being (physical, material, social and emotional). Autonomy, activity, material indexes, economic resources, health, living conditions, intimacy, safety, place in community and personal relationships are also key indicators in assessing the quality of life of an elder. Fernandez-Ballesteros (1998) in [17], gives an insight into the measurement of quality of life by the context and circumstances in which the elderly lives, such as social status, age or gender, although corroborating the position of several authors, stating that the generalized standard patterns of quality of life can not be established. In sum, it seems generally agreeable that seniors quality of life is directly related to biological, psychological, social and behavioral factors, and only through mediation of all these factors we can think in standards patterns of quality of life for seniors.

Chapter3

Knowledge Discovery and Data Mining

Nowadays companies and institutions from various areas, such as science, business, or health care around the world have a huge amount of data stored with varied information, where the increase in the degree of complexity in their structures have a exponential tendency. The emergence of differentiated technologies for storage and retrieval of data, allow the user or analyst to obtain strategic information that can be transformed into useful knowledge from a collection of a logical structured data.

3.1 Knowledge-discovery

Worldwide companies and organizations have big databases aggregated to their computer systems. The analysis of such voluminous amount of data is not humanly possible without the assistance of computational tools. In this context, it is important to understand all the terms and the differences between *Knowledge Discovery* (KD), *Knowledge Discovery in Databases* (KDD), *Knowledge Discovery and Data Mining* (KDDM) and the specific step of *Data Mining* (DM).

Knowledge Discovery (KD) is a process through wich new knowledge or important information about the study domain is acquired. It involves many steps and each step consists of a particular discovery task [28].

Knowledge Discovery in Databases (KDD) regards the application of the KD process to databases [28].

Fayyad et al. [16] defines it further as the non-trivial process of identifying valid, novel, potentially useful and ultimately understandable patterns in data. These result from the intersection of several areas, such as Machine Learning, Pattern Recognition, Databases, Statistics, Artificial Intelligence, Knowledge Acquisition for Expert Systems, Data Visualization and High-performance Computing.

The Knowledge Discovery and Data Mining (KDDM) process regards to the *KD* process applied to any data source. It is considered the entire knowledge extraction process, which, according to [28], addresses issues like data storage and access, algorithms efficiency, results interpretation and visualization and, human-machine interaction.

Data Mining and *Machine Learning* are part of all these processes. In the following subsections, we contextualize them in the knowledge discovery process.

3.1.1 Data Mining

Data-Mining is a particular step of the KDD or KDDM processes. It consists on the application of specific algorithms and a variety of analysis tools to extract patterns and relationships from data that can be used to make valid predictions. It component of KDD currently strongly depends on known techniques such as Machine Learning, Pattern Recognition and Statistics to find patterns in data. However, it has become better known than the KDD and KDDM process themselves mostly due to it being the step where the search of knowledge techniques are applied. According to [28] and depending on the goal this knowledge discovery is distinguished into two types: *Verification* and *Discovery*. *Verification* only verifies the user's hypothesis. *Discovery* is related to automatically finding new patterns. It can be subdivided into *Prediction*, where the found patterns are used to predict the future behavior of some entity; and *Description*, where the found patterns describe some entity in a way which is comprehensible to humans. The algorithms or methods in this task are supervised by an external source who knows the associated output values for each set of input attributes. According to the target variable type, numeric or categorical, we have *Regression* or *Classification* techniques. The algorithms used in *Description* assume the inexistence of a target variable, following here the system find patterns for presentation to a user in a way which is comprehensible to humans. The algorithms used in the this task ignore the output attribute following the unsupervised learning paradigm where is the *Clustering* and among others techniques. Figure 3.1 shows the taxonomy we have just described. However, we must stress that this a very elementary and over-simplistic view of Data Mining.

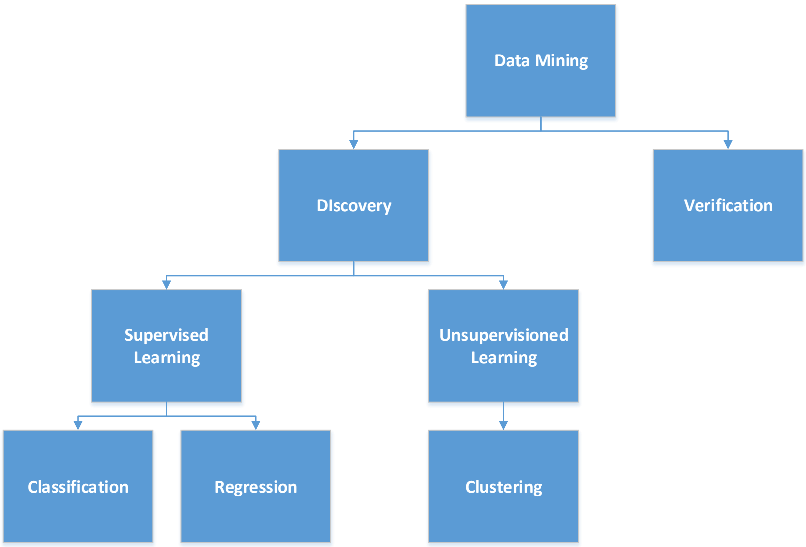


Figure 3.1: Data Mining Taxonomy.

3.1.2 Machine Learning

Along with research fields, such as statistics and pattern recognition, Machine Learning is one of the research field on which Data Mining relies on. It concerns the design and development of learning algorithms that allow computers to automatically learn patterns or make decision based on data. To analyze the data is necessary an algorithm, which is a sequence of instructions that must carry out the processing of an input as an output. Although in some cases there is no algorithm to perform the desired task, it is only known the input and the output as it should be, not knowing how to do the transformation of the input in the output. This lack of algorithmic knowledge is compensated in the amount of data stored where the automatic extraction of knowledge from this data is the key to achieve the desired output by the analyst.

In 1959, Arthur Samuel defined the area of Machine Learning as a field of study that gives computers the ability to learn without being explicitly programmed[48]. This learning is achieved through the collection of data, direct experience or instructions that may be helpful in completing a task or make predictions. It is a subfield of Artificial Intelligence that consists of learning how to better perform given tasks in the future, based on past experience, always looking for automated methods without human intervention or assistance. In [35], Tom Mitchell states that a computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks T as measured by P improves with experience E.

For there to be complementarity between learning and experience, Machine Learning intersects the area of Computer Science with the field of Statistics [36]. On one hand, Computer Science seeks the algorithmic efficiency for solving the optimization problem, processing, and storage of data, and on the other hand, Statistics focuses mainly on conclusions that can be inferred from data by constructing mathematical models. In certain applications, the efficiency of inference (like space complexity and time) can be as important as the accuracy of prognosis.

Much of the progress in Machine Learning is due to the high compatibility of its applications in various areas in the real world. **Web page ranking** is used in search engines when a query is submitted with the desired search, by returning the number of answers in order of their degree of relevance. **Speech Recognition** is a program that is currently available in all commercial systems. The accuracy of this system is higher since the beginning of the automatic learning system instead of the manual system [36]. *Robot Control* methods have been successfully implemented in robotic systems, such as in learning precision aerial maneuvers of airplanes [36]. **Computer Vision** such as face recognition and microscopic cell classification systems, use automated methods for recognition [36]. **Medical Diagnosis** algorithms, such as Decision Tree Learning, Naive Bayesian Classifier, and Neural Networks, are currently used for diagnosis of patients suffering from a disease [27].

3.2 Data Mining Process: Cross-Industry Standard Process

The Cross-Industry Standard Process for Data Mining (CRISP-DM) is an industry-proven way that offers a well-defined structure for the knowledge discovery process. The goal of proposed structure is to achieve better and faster results in Data Mining projects. This process was proposed in 1996 by a consortium of four companies, SPSS (a provider of commercial Data Mining solutions), NCR (a database provider), Daimler Chrysler, and OHRA (an insurance company)[28]. This consortium, formed with the goal of developing CRISP-DM, requested several Data Mining practitioners to develop a standard model to serve the Data Mining community. After several improvements and reviews, CRISP-DM was proposed by *Cios et al* in 2000, resulting in a process with six phases and intended to be a comprehensive Data Mining methodology and process model that provides a complete model for conducting a Data Mining project[28]. This process has been used in several projects from distinct areas such as, analysis of thrombosis data, text mining, analysis of retail store data, among others[28]. The phases that constitute this process are *Business Understanding*, *Data Understanding*, *Data Preparation*, *Modeling*, *Evaluation* and *Deployment*, as shown in Figure 3.2.

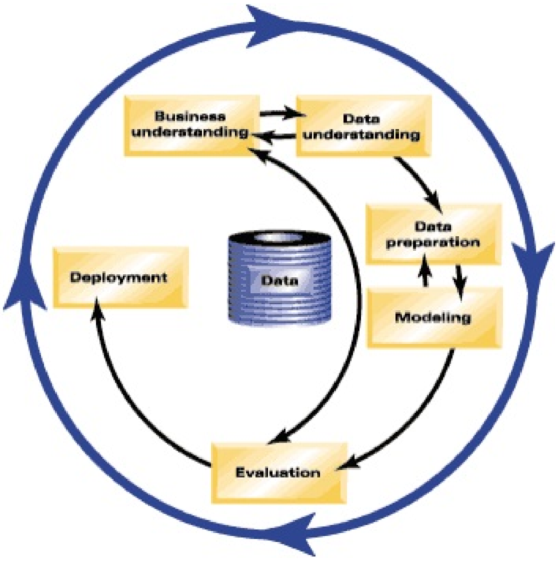


Figure 3.2: Process diagram showing the relationship between the different phases of CRISP-DM (reprinted from [54]).

In the following subsections we present each of these six phases [54]:

3.2.1 Phase 1: Business Understanding

Probably the most important phase of any Data Mining project, is focusing on understanding the projects objectives, transforming into a formal Data Mining definition, consequently developing a primary plan designed to achieve the projects goals. It is necessary to determine the business's objective, and all the clients goals need to be clear and describe all the resources available to reach these goals.

3.2.2 Phase 2: Data Understanding

This step starts with an initial data collection, where the analyst has to explore the data to clarify all the complexity, identifying data quality problems or detecting interesting information, such as data subsets, to form hypotheses about hidden knowledge and build the final data from the original data.

3.2.3 Phase 3: Data Preparation

This phase has the purpose of building the final data from the original data to use with modeling tools, where are included tasks such as, records, tables, format data, cleaning of data and construct new data.

3.2.4 Phase 4: Modeling

Several modeling techniques are selected to calibrate and apply Data Mining methods to the prepared data, but some techniques have specific requirements on the data format. In this phase it will be probably necessary to step back to the Data Preparation phase. Modeling steps include selecting the modeling technique, choosing the right algorithm, such as, for example, decision trees or neural networks. The next step is to generate a test design to determine the reliability of the model. For supervised learning the test is usual made by quality measure, developing a model based on the set of existing data and test the validity using a separate set of data. For unsupervised learning, such as the clustering technique, measurements may include criteria such as ease of interpretation, deployment, or required processing time.

3.2.5 Phase 5: Evaluation

It is necessary to evaluate the generated knowledge from a business perspective after the final deployment of the model is built by the data analyst. At this stage it is critical to determine if any important business issue has not been sufficiently analysed.

3.2.6 Phase 6: Deployment

The obtained knowledge must be organized and projected in a customer-oriented way, which often involves applying real-time personalization's, such as web pages. It is necessary to implement a strategy to deploy all the Data Mining results in the best way, produce a monitorization plan to avoid incorrect usage of Data Mining results and produce a product final report.

Chapter4

Market analysis and Existing Solutions

In the recent years there has been a significant growth in applications related to the health sector, in which the user has at its disposal several solutions for analysis and control over health-related behaviors, such as diabetes, physical exercise and medication control. Surprisingly, little attention is still given to depression as there are few tools to support detection and treatment of mental disorders such as depression.

Most of the developed systems with the aim of predicting depressive problems have a failure in the specification, because they are based on general risk factors in the prediction of depression in the life of an individual. The lack of using personal data, which is critical for the specificity and reliability of the study of depression of the individual, such as physical activity, the analysis of the trend in the social community and its usual location, are used by few systems on the market and its availability for a personal use is still limited. These same applications present low realization of their main goals, that is, they are slightly explanatory about causes and factors influencing the onset of depression. It is important to notice that as far as we know through the specifications of this tools, only uses Domain Knowledge instead of Machine Learning or Data Mining techniques.

Next, we examine some of the solutions created in order to monitor and prevent some depressive symptoms:

Mobilyze application was developed by researchers at Northwestern University (USA) and the main goal is monitoring behavioral patterns and mood states by identifying states that trigger depression, thus preventing it. The strategy consists of a combination between the sensors of personal phone (such as GPS, accelerometer, and Wi-Fi), with the data provided by users (such as the state of mind and social context). The main advantage of this application focuses on the possibility of anticipating depressive moments of the individual being monitored. Still, there are some barriers that have not been exceeded, for example, the differentiation of a calm day to a sign of depressive disorder[33].

Xpression is a mood analyzer, developed by the British company Ei Technologies, that enables monitoring the state of mind of the individual using only the voice, more specifically, through the attribution of emotions during calls made by the user in their day-to-day, thus analysing the state of anxiety, stress and depression. This application is not available in the distribution market , although it is recognized it is importance in studies conducted in this field[46].

Menthal was developed by researchers at the University of Boon, in Germany, for the Android operating system. The main purpose of this application is to monitor the use of the personal phone, recording the time that the user spends on the phone, analysing which applications are used more often. This data is sent to an anonymous server where the information and statistics of each user are collected, and then analyzed by experts. Studies are being made using this application, with the aim of developing a component which allows for detecting depression[47].

Another type of existing solutions aim to guide or help the user with a tutorial to prevent depressive disorders, based on general risk factors of depression. In most of these applications, the user is required to answer the inquiry based on established scales for monitoring this disorder, so there are not any specificity of data on the user's phone to monitor depression.
We now list some solutions in the market following these guidelines:

CBT Depression Self-Help Guide was developed by the Limited Liability Company, Excel at Life (USA). It is a guide support to detect depression, providing therapeutic articles, diary of cognitive thoughts to learn to challenge thoughts causing stress, providing positive thoughts, suggestions for helping track the concentration through motivation and screening tests with graphic support to monitor the severity of depression[15].

MyM3 is an application developed by collaborators at the University of Georgetown, Washington (USA) that, in collaboration with cognitive behavior therapists have created a list for self-evaluation of primary care to monitor potential symptoms of anxiety, indicating the relative risk of depressive symptoms or traumatic disorders. Such evaluation is made by the user and the data is then sent to a designated health care professional that analyzes the questionnaire and obtains relevant information helping the psychologist on getting a better diagnosis at the first visit with the patient[38].

Depression Calculator is an application developed by Egton Medical Information System Limited (UK), based on Patient Health Questionnaire Scale (PHQ-9), through which the user responds to this standard questionnaire trying to obtain a diagnosis of the depressive state. It is also provided a digital informative leaflet with important information and advices about Depression and antidepressants previously analyzed by experienced authorities in the field[40].

Chapter5

DepSigns: A CRISP-DM approach to Depression Signs Detection through Smartphone Usage Data

Throughout this chapter are presented the six phases of the DepSigns project, following the CRISP-DM process.

5.1 Business Understanding

Probably one of the most important phases of any Data Mining project is focusing on understanding the project's objectives, transforming into a formal Data Mining definition, consequently developing a primary plan designed to achieve the project's goals. It is necessary to determine the business' objective, and the client's goals and needs to be clear and describe all the resources available to reach these goals.

Our project specification required the use Machine Learning to infer depressive symptoms in elders. Psychologists gave us the information that depression detection is mostly related to changes in habits. Therefore, we decided to also conduct a statistical analysis on elder's data to cover this issue. However, the statistical analysis soon proved to have issues, such as when the elder is already depressed when we first start collecting data.

The Machine Learning process is based upon the psychologist's manual process, which is described in Section 5.2. It specifies which data should be analysed and how we decided on the structure, layout, and organization for our project, and chose an algorithm which would suit those requirements. We believe our method is closely related with the psychologist's manual process and, therefore, results should be similar, but only further experiments will tell.

Most of this process consisted of a set of requirements from our project specification and, therefore, a deep analysis on the business understanding for this project is outside the scope of this thesis.

5.2 Data Understanding

This project handles sensitive data, which is personal to seniors. Therefore, there are some restrictions regarding how data can be disseminated. For this reason, we have created a data generator which simulates the physical and psychological states of abstract individuals. Next, we describe how this data is stored, organized, and how we have preprocessed it to build our dataset.

5.2.1 Data Generation

At this stage, it is important to mention that no data source whatsoever was provided for this project. Originally, all the data was intended to be originated from the Smart Companion project of Fraunhofer Portugal, which is an android customization that was designed to address senior's goals and needs, with the goal to support seniors in their daily activities[41]. However, the database did not fulfil the requirements for this project and there was not enough data collected to support a study of this kind. For these reasons, we decided to create a generator which simulates the necessary data. This has imposed a great challenge on the replication of real-life situations. Using computer generated data makes it impossible to guarantee that the database represents actual real-life situations, as there is no way of matching this data to real-life everyday cases. As such, the generator was created under psychologist guidance, which have provided feedback on common cases on which data would be useful, as well as its relevance in this study. The generator uses a probabilistic approach to this problem. Instead of having data from real patients, we have generated a list of patients, as well as their behavioural patterns for the duration of several weeks. This was achieved by using a *Beta* distribution on the moods of the patients, from which the rest of the data is inferred.

Beta distribution is a family of continuous probability distributions defined on the interval $[0, 1]$ configured by α and β shape parameters, which are the exponents of the random variable, and control the shape of the distribution. This distribution assumes the modulation of the behaviour of random variables limited to intervals of finite length. There are several areas of statistical description that use this distribution, such as genetic population or genetic heterogeneity in the probability of HIV transmission, among others. The standard *Beta* distribution gives the probability density of a value x on the interval $[0, 1]$, as shown in Equation 5.1.

$$P(x) = \frac{(1-x)^{\beta-1}x^{\alpha-1}}{B(\alpha, \beta)} \tag{5.1}$$

where:

$P(x)$ is the probability function

α, β are the distribution parameters

$B(\alpha,\beta)$ is the *beta* function

Figure 5.1 shows a few examples of possible *Beta* distributions.

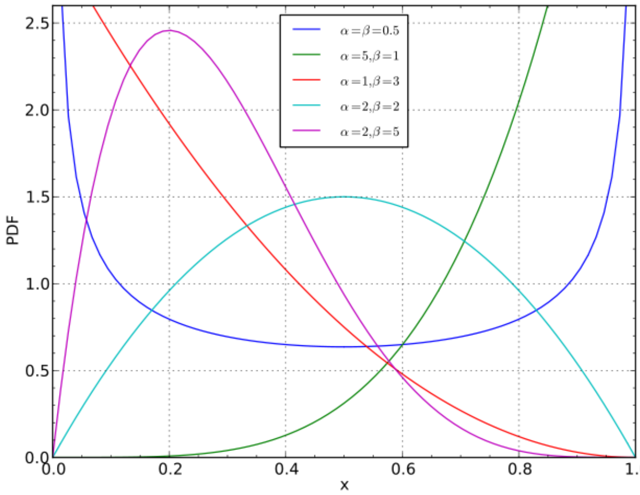


Figure 5.1: Probability density function (reprinted from [3]).

From the beginning, this was intended to be a preliminary study. It is not known to us that any previous study of this kind exists, which limits to a large degree the basis we had as a starting point. Having no background on how to automatically detect depressive symptoms using our methodology, we chose our methods based on psychologist feedback and guidance, providing that our methods would closely relate to the methods used by the psychologist. For the psychologist analysis, it is important to know:

- how did the elder perform, activity wise?
- how did the elder perform, location wise?
- how did the elder perform, communication wise?
- and so on.

The order in which these fields are evaluated is relevant. This means that the data concerned with elder's activities is considered the most relevant by the psychologist, under which a failure would indicate immediate depressive symptoms. The order of relevance given by the psychologist for each field is the following:

1. Activities
2. Locations
3. Communications

    1. Calls
    2. Not Answered Calls
    3. Messages

4. Moods
5. Ludic Activities

The statistical study shows as a complement to the Machine Learning process in two distinct ways:

- it complements the study by giving more relevance to personal habits;
- it allows for a much better data visualization tool.

We decided on using a generic central tendency metric which would allow future work to extend the set of available methods. On top of that, the application is also able to adapt to different methods used by different psychologists, provided that a suitable evaluation function exists and is implemented. However, we did focus on analysing the last two weeks of data, as previous studies exist that indicate that two weeks are sufficiently relevant to provide reliable results, as defined by the *American Psychiatric Association* [5]. One last thing to consider, related to the statistical process, consists in defining how the data is first evaluated. As an example, we consider the number of calls and not the time spent talking on the phone. The following lists how we evaluated each field as well as the reason that justifies our approach:

- **Activities** - total time spent performing a given activity. The amount of time is considered relevant as it yields a better indicator of daily physical activity then a count of the number of activities performed. We consider nightly activities to be prejudicial to the individual, as it indicates poor sleeping habits.
- **Locations** - total time spent away from home. Just leaving the house is not indicative of a positive activity, as there is no indication of the kind of activity. For instance, several activities could be performed, out of the house, like buying bread, which is not as relevant as leaving the household for several hours. Even so, we have no way of identifying the quality of the absence from home. That is, leaving the house, for how long as it might be, does not indicate whether the

visited place contributes positively to the elder's state of mind, and there is no automatic process, that we are aware of, that does so. As such, we consider all locations to have equal relevance, and make no attempts on identifying the location whatsoever.

- **Communications** - number of calls or messages sent and/or received. The amount of time spent using the phone was considered irrelevant by the psychologist, as the amount of time does not vouch for the quality of what is being talked; instead, we only count the number of actions taken using the phone. Take as an example a situation in which an elder spends several hours talking to the same person. This could either indicate a very active social life or a very strong need for attention.

- **Moods** - count of the weekly mood self-evaluation. Seniors have four options to self-evaluate: *Bad*, *Not Well*, *Fine*, and *Very Good*. Each mood symbol is given a value by assigning to it a position in an array. Three is then subtracted from this position, making sure that bad moods yield negative values and good moods positive ones. The sum of these values is used as the the weight factor for each weekly mood.

- **Ludic Activities** - amount of time spent performing the activity weighted by the achieved score. The reason for this relates to elderly cognitive changes, as detailed in Chapter 2, Section 2.1.

For each name from a list of hand-given names the generator creates a patient entry. Each patient is then given a mood which we generate by assigning random integer values between 1 and 5 for the α and β parameters of the Beta distribution function, respectively. The reason these values are random is mainly because it is unknown to us what the correct distribution would be, thus allowing us to conduct several studies and getting to some final result which would seem reasonable. Again, the psychologist feedback was essential at this point, which has later confirmed reasonability after inspecting several cases. Once again, this is only justifiable from the absence of any real-life data source.

As the probabilistic function generates real values between 0 and 1, we then assigned thresholds for mood values. According to psychologist advice, the moods can be divided in four categories. Each category was given a 25 percentile of the range provided by the distribution function. As such, any mood value below 0.25 would be considered *Bad*, between 0.25 and 0.50 as *Not Well*, between 0.50 and 0.75 as *Fine*, and above 0.75 *Very Good*. These categories are then used as thresholds in the generation of all other data referring to the same patient, in the same week. One mood value was generated per week for the entire period of 12 weeks.

Every other field is generated by using the same Beta distribution parameters; that is, the α and β values that were randomly generated for the mood. However, the mood plays an important role in the generation of these fields, as it is used as a threshold for whether the patient complies or not with some obligations. Because the Beta distribution generate good habits for bad moods, and vice-versa, the worst the patient's mood the more we would require of him/her. That is, for values generated between 0 and 1, indicating how willingly the patient would comply with his/hers obligations.By *obligations* we mean any average person's daily duties, like answering phone calls, going out, waking up in the morning, among others. We assume that a patient having a *Bad* mood would require a willingness threshold of 0.90; however, we would require just 0.70 for patients with a *Not Well* mood; 0.50 for *Fine*; and 0.25 for *Very Good*. Again, these thresholds are merely experimental and no sustainable data will ever be available unless provided from real-life situations.

5.2.2 Database Model

The database was modeled to represent a set of constraints that were agreed with psychologist guidance. The goal, however, was to match, as closely as possible, the structure of the Smart Companion database, without compromising the physiologist's requirements on elder data. To achieve this, the number of fields present in the database was kept at a minimum, allowing for improvement when trying to match it with the Smart Companion database.

Several entities were listed by the psychologist as essential for this study (see Figure 5.2):

- **Users** are a presentation of a studied entity, which, in the context of the current application, means an elder. This field is central to the database and is used to relate all other fields by means of *foreign keys*. The data stored in this table encompasses the name and date of birth for each elder.
- **Activities** represent daily physical activities separated by three different periods of the day: morning, afternoon, and night. The time at which the activity takes place indicates the period to which it belongs. These periods are later used to influence the relevance of the activity in the elder's schedule. Each entry consists of a start and end timestamps indicating the period of the activity.
- **Locations** indicate places that seniors visit when they leave home. The place itself is not stored, instead just the period in which the person was absent from their place of residence. This decision was justified by the computer's inability to decide whether any place was a good or a bad influence on any one given person. As such, just the timestamps for when the elder leaves and later arrives at home were stored.
- **Phone calls** are a registry of the elder's voice communications over the telephone. Again, because the concrete context of the call cannot be determined as being positive or negative for the person under study, the only fields that were stored where the timestamps for the beginning and end of each phone call. This also allow us to compute the amount of time spent on the telephone.
- **Not Answered Phone Calls** consist of a registry of missed phone calls, not answered by the peer under study. Because of our inability to know the relevance of the other peer for the elder's mood, we only store timestamps. In this case, however, and because missed phone calls do not have a duration period, we only register one timestamp, indicating the time at which the call was missed.
- **Messages** could be text messages or messages of other kinds of media. Again, the inability to infer whether the other peer is either a positive or negative influence to the person under study, forced us to discard any information about that peer, keeping the database structure minimal. For this field only one timestamp is stored, indicating when the message was first received.
- **Moods** are supposed to be retrieved from elder feedback. That is, the senior him-

self/herself would provide daily feedback on how good his/hers mood is. However, because a generator was used, this field is randomized as described in Section 5.3. Therefore, ideally, this field would be an introspective representatition of the elder's mood in a daily basis. This field stores one of four possible different mood values (*Very Good*, *Fine*, *Not Well*, or *Bad*) as well as a date indicating when the feedback was given.

- **Ludic Activities** are a registry of conginitive activities performed by the senior which hold a score. The context or content of the activity is not known, but the score is considered relevant (that is, seniors with no depression symptoms are expected to achieve better scores). For these reasons, only the start and end timestamps for the activity, as well as the performance score are saved in the database.

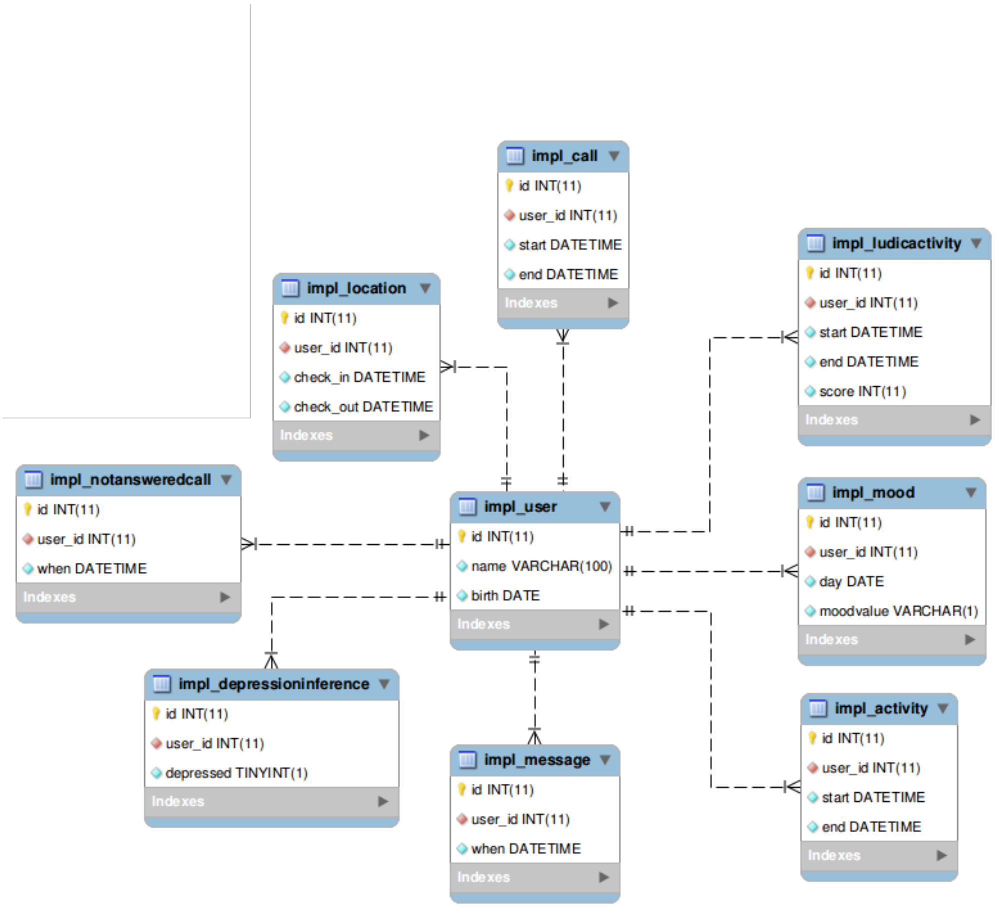


Figure 5.2: The database EER model

5.2.3 Statistical Analysis

We conduct a statistical analysis on this data and use it to infer depression symptoms from personal habits. The goal of this statistical analysis is to find deviations from the elder's usual behaviour, by detecting standard patterns in the elder's daily activities, and causing field-wise alerts to be shown to the psychologist or caregiver. It is not clear to us - or to anyone, as far as we know - which kind of analysis would best detect the mentioned deviations. For this reason, we use an abstract evaluation function which can be configured by the psychologist using this software. Also, we provide a few samples which we consider to be descriptive of the application's capabilities, although these routines are in no way intended to be used as a final product. Using such an abstraction allows the psychologist to further study which kind of analysis would be best, by changing the central tendency metric and watching the application respond, in real time, by changing conclusions on the elder's conditions. The set of evaluation functions is further expandable, meaning that it was our intent from the beginning not to create a closure on the set of provided methods. However, as of the moment of this writing, adding evaluation functions requires making changes to the source code.

5.2.4 Central tendency measures

The evaluation function is considered to yield a measure of centrality of some sort. There are no restrictions on this measure, although it is highly recommended that the routine relates to the data as closely as possible. As input, we provide at each iteration an array with the data that has been processed by the previous iterations, plus one entry which has not yet been processed. To achieve this effect, while we iterate through the data samples, ordered by week, we keep an array containing all previous data entries, as shown in Listing 5.1. The output will later be used as a threshold for the weeks to come; we project this concept in the graph using splines, as shown in Figure 5.3. The result of the evaluation function is kept in a second array, which holds values for spline lines. Later, these lines are compared to the elder's weekly results, providing a threshold on how well the elder is doing. This threshold is considered to be the minimum requirement for us to consider the individual's week as fulfilled. In case this mark has not been met for the last two weeks in a row, then an alert is raised for the given field. The decision to consider the last two weeks alone was achieved under psychologist guidance. Past weeks are used to infer personal habits, so we can make conclusions on the elder's current state.

```
splinevalues = []
previous = []
for week in data:
    spline = Evaluate(week,previous)
    splinevalues.append(spline)
    previous.append(week)
```

Listing 5.1: Algorithm describing the process of collecting spline values using the generic evaluation function.

Figure 5.3 shows a sample projection of the data and statistical analysis. Weeks are listed in the abscissa axis, and the ordinate axis lists compliance. Each vertical bar indicates how well a elder as performed during the matching week. The spline indicates a measure of centrality, according to the chosen evaluation function. Bars below the spline indicate a poor performance for the given week, which could indicate depressive symptoms.

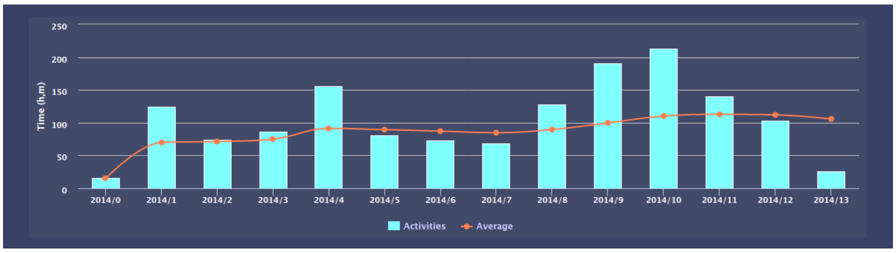
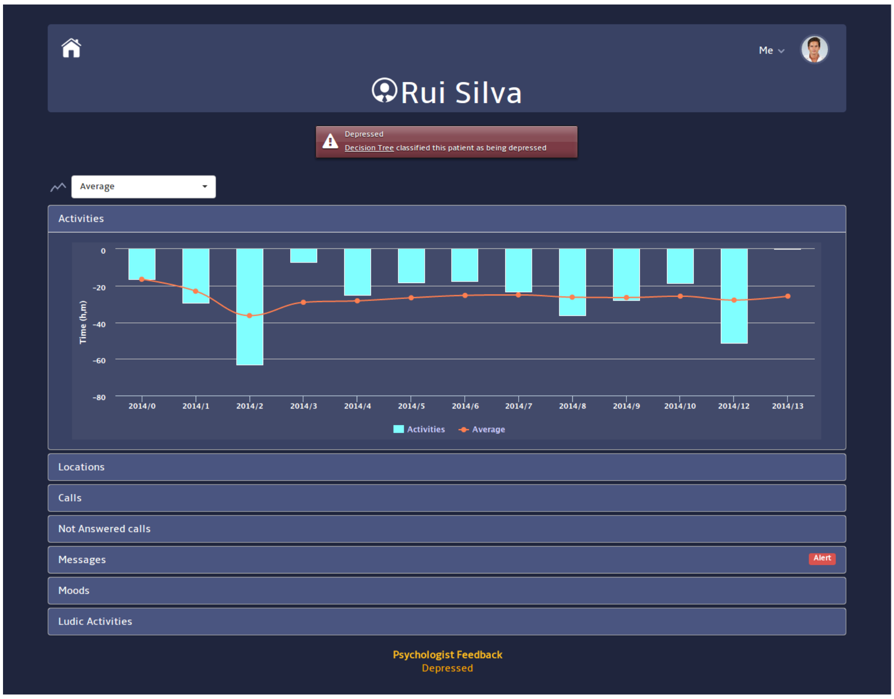


Figure 5.3: Sample projection of the data and statistical analysis.

Nevertheless, an individual analysis such as the one given by any of the centrality metrics may be misleading as it heavily dependent on the senior's initial condition. Elders that already performed poorly when we first started sampling data will accuse as being in their normal condition if they continue performing the same way, as shown in Figure 5.4.

Therefore, we will not be able to identify any depression symptoms whatsoever. This is why we also conduct a study using Machine Learning, which tries to solve this problem.



Depressed
Decision Tree classified this patient as being depressed

Psychologist Feedback
Depressed

Figure 5.4: Example of an individual's weekly time of activity, as well as its corresponding average

This kind of data analysis is therefore used only to find deviances from personal habits, whatever those habits might be. Because we are well aware that conclusions achieved this way are provisional, we project the data history and allow the psychologist to further study each case individually. This history (shown previously in Figures 5.3 and 5.4) projects the data comparing it to the threshold generated by the generic evaluation function, providing a visualization tool for what is happening in the background . Later, we will see how this can also be used as a mechanism to provide feedback to the server, so we can infer conclusions from the general population.

The central tendency measure, therefore, plays a central role in the process of finding deviances to the typical behaviour. Because this routine provides the threshold for comparing against the data, it consists of an important backbone of this application. However, and because it is not known what the best measure would be, we have defined four different centrality measures that allow studying how well the application would perform under different conditions. These routines were suggested by the psychologist, having both parts agreed that they would in no way be final:

- average;
- simple moving average using 3 week intervals;
- simple moving average using 5 week intervals;
- standard deviation.

Each of these methods generates a threshold value that relates to the data in some way. Also, each will yield different results. It is now up to the psychologist to understand how well these results relate to each individual's condition and, possibly, try to infer a general case.

Although *average* may refer to several different concepts of centrality (such as *geometric mean*, or *harmonic mean*), throughout this text we will always refer to *arithmetic mean* unless otherwise specified. As such, we define *average* as being a measure of central tendency. The average, usually represented by $\overline{X}$, is defined as the sum of all the values in a data set divided by the number of values in the sample [49], as described by the following arithmetic formula:

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \tag{5.2}$$

where:

n is the number of observations

X_i is the i-th observation of sample X

In Statistics, the standard deviation, usually represented by the Greek letter σ, is the most common measure of statistical dispersion, indicating how much variation or dispersion there is from the average. A small standard deviation indicates that the data tends to be tightly related to the average. Data sets with large standard deviation indicate that the data is spread over a wider range of values[2]. The following formula represents the mathematical standard calculation:

$$\sigma = \sqrt{\frac{\sum(X_i - \overline{X})^2}{n}} \tag{5.3}$$

where:

X_i is the i-th observation of sample X

$\overline{X}$ is the average of all values of sample X.

n is the number of observations

Weighted moving average is a statistical method for smoothing time series by averaging with weights to a fixed number of consecutive terms. This method is commonly used with time series data to smooth out short-therm fluctuations and highlight longer-term trends. Given a series of numbers and a fixed subset size, the first element of the moving average is obtained by taking off the initial fixed subset. Then the subset is modified by shifting forward, excluding the first number of the series and including the next number, creating a new subset of numbers, which is averaged[13]. This is represented by the following formula:

$$\overline{X}_w = \frac{\sum_{t=1}^{n} W_t * V_t}{\sum_{t=1}^{n} W_t} \tag{5.4}$$

where:

t is the index of the entries of an ordered set of samples

V_t is the sample with index t from the set of samples V

W_t is the weight factor for the sample V_t

n is the number of periods in the weighting group

5.3 Data Preparation

Our generator and the Smart Companion database are interchangeable. This does not mean that they are the same thing, but rather that one is to be used in replacement of the other. In this case, the generator was created with the sole purpose of serving as a temporary replacement of the Smart Companion database, due to its unavailability (see Chapter 5, Section 5.2 for more details).

A very small set of queries was constructed which would ease further access to the data, and even abstract the process. These queries have the following two purposes:

1. **Data Reduction** - The raw data is stored in a daily basis. However, for the purpose of this study, and under psychologist guidance, it was considered irrelevant to go as deep as analysing daily data. According to the psychologist's opinion, the conclusions become more sustainable when analysing weekly periods, and therefore the first step consists in aggregating the data in that way.

2. **Data Transformation** - After being aggregated, we create weighted resumes of the data. How each resume is computed depends on the actual field being weighted. We next proceed to explain this process.

As far as Data Transformation is concerned, the set of fields in the database can be divided in 5 different groups.

- Compute three summations of the time spent performing an activity divided by periods of the day (morning, afternoon, and night, depending on the time of the day). Each summation is then multiplied by a constant weight factor which as been specifically assigned to its matching period. In this case we used a multiplier of 2.0 for the morning period, 1.0 for the afternoon, and -2.0 for the night period. These values, however, are not justified and should be considered future work. For this reason, we left the constants out of the query allowing for easy customization. Then all values are added up. This way we give more or less relevance to each period of the day, while still ending up with a single, descriptive, value. Only the field for *Activities* fit into this category.

```
SELECT id, SUM(period * duration) as weight, week
    FROM(
        SELECT id, TRUNCATE(SEC_TO_TIME(SUM(TIME_TO_SEC(TIMEDIFF(end,
            start)))) /60,0) AS duration,
            (CASE
            WHEN TIME(start) BETWEEN '06:00:00' AND '14:59:59' THEN %f
            WHEN TIME(start) BETWEEN '15:00:00' AND '23:59:59' THEN %f
            WHEN TIME(start) BETWEEN '00:00:00' AND '05:59:59' THEN %f
            END) AS period,
            CONCAT(YEAR(start),'/',WEEK(start)) AS week
        FROM impl_activity
        WHERE user_id = %d GROUP BY week,period         ORDER BY start
    ) AS tmp
GROUP BY week
```

Listing 5.2: Query used for Activities.

The time periods are given as arguments to the query (shown as *%f*) yielding the weights used for each period. There are three daily periods (morning, between 6am and 3pm; afternoon, between 3pm and midnight; and night, between midnight and 6am). The week for each entry is represented in the format YEAR/WEEK. The user ID is also a parameter represented here as *%d*.

- Compute the summation of time spent performing the activity on any given week multiplied by the performing scores. This gives not only the amount of time spent performing those activities, but also the weighted quality of the result. Only the field of *Ludic Activities* fit into this category.

```
SELECT id, week, duration * score AS weight
FROM (
  SELECT
    id, score,
    TRUNCATE(SEC_TO_TIME(SUM(TIME_TO_SEC(TIMEDIFF(end,start)))) / 60, 0)
         AS duration,
    CONCAT(YEAR(start),'/', WEEK(start)) AS week
  FROM impl_ludicactivity
    WHERE user_id = %d
    GROUP BY week
    ORDER BY start
  ) AS tmp
```

Listing 5.3: Query used for Ludic Activities.

The score will be used as a weight factor to leverage the quality of the result.

- Compute the summation of weighted mood values for each weekly period. Each mood is assigned a different weight, being -2 *Bad*, -1 *Not Well*, 1 *Fine*, and 2 *Very Well*. These values are not relevant *per se*, as long as bad moods are negative and good moods are positive, and equally relevant states of mood are equality distant from zero (that is, *Bad* is worth -2, while *Very Well* is worth it is absolute value). This way the summation of these values yields the overall quality of the elder's mood state. Only *Moods* fit into this category.

```
SELECT id, CONCAT(YEAR(day), ’/’, WEEK(day)) as week, SUM((FIND_IN_SET(
   moodvalue, ’B,N,?,F,V’) - 3)) AS weight
FROM impl_mood
  WHERE user_id = %d
  GROUP BY week
```

Listing 5.4: Query used for Moods

Each mood symbol is given a value by assigning to it a position in an array. Three is then subtracted from this position, making sure that bad moods yield negative values and good moods positive ones. The sum of these values is used as the the weight factor for each weekly mood.

- Compute the total amount of time spent performing a given obligation in a given week. This group applies to *Locations*.

```
SELECT id, TRUNCATE(TIME_TO_SEC(SEC_TO_TIME(SUM(TIME_TO_SEC(TIMEDIFF(
   TIME(check_out), TIME(check_in)))))) / 60, 0) AS weight, CONCAT(YEAR
   (check_in), ’/’, WEEK(check_in)) as week
FROM impl_location
  WHERE user_id = %d
  GROUP BY week
```

Listing 5.5: Query used for Locations

The time spent by the elder out of home is computed as the difference of the check-in and check-out timestampos, in total number of minutes.

- Count how many times a given obligation was fulfilled in a weekly period. This group applies to *Calls*, *Not Answered Calls*, and *Messages*.

```
SELECT id, CONCAT(YEAR(impl_notansweredcall.when), ’/’, WEEK(
   impl_notansweredcall.when)) as week, COUNT(*) as weight
FROM impl_notansweredcall
  WHERE user_id = %d
  GROUP BY week
```

Listing 5.6: Query used for Not Answered Calls, although Calls and Messages use a similar query.

One of the consequences of this process of Data Transformation consists in reducing every weekly data set to a single entry, which consists of a numeric, weighted, representation of the senior's performance on any given week.

A training sample consists of a matrix with N rows (observations $O_i, 1 \leq i \leq N$) and M variables $x_j, 1 \leq j \leq M$. There are exactly K classes, which must be known before hand, and which are used to classify the observations. In our case, there are only two classes, as we categorize elders as either having depressive symptoms or not. The number of rows N is dictated by how many different elders the psychologist has given feedback about, and the number M of variables is fixed and corresponds to twice the number of fields we are currently studying, as each field is analysed for the past two (2w) and four weeks (4w). Table 5.1 shows a sample of all fields discussed in Section 5.2. Each row is analysed by the CART algorithm to generate the Decision Tree, and the Classification column is used to classify the trained data.

Observation	O_1	O_2	...	O_N
Activities (2 weeks)				
Activities (4 weeks)				
Locations (2 weeks)				
Locations (4 weeks)				
Calls (2 weeks)				
Calls (4 weeks)				
Not Answered Calls (2 weeks)				
Not Answered Calls (4 weeks)				
Messages (2 weeks)				
Messages (4 weeks)				
Mood (2 weeks)				
Mood (4 weeks)				
Ludic Activities (2 weeks)				
Ludic Activities (4 weeks)				
Classification				

Table 5.1: Example of a training sample.

5.4 Modeling

We use a Decision Tree to classify depressive symptoms, mostly due to its high level of interpretability. To generate this tree, we allow the psychologist to provide feedback on a subset of elders, indicating whether depressive symptoms exist or not. We commit this feedback, in the form of a boolean flag, to the database associating it with the given elder instance. We can then generate a training set by correlating the elder's history with the provided feedback. We generate this set in a format that the scikit-learn framework [9] understands and is capable of generating the Decision Tree from. We then save the tree to a file using *pickle* [22], enabling us to load it later, on demand, without having to generate it again. We also generate PDF and SVG files with the rendered tree, which is useful as visual support. We use the same format as the training set to classify other samples not used as training. Therefore, we generate a table like we did for the training sample, and use it to predict the elder's state of mind according to previously trained rules. Contrary to the statistical analysis, the decision tree classifies symptoms according to global population standards. This study is useful in identifying elders with depressive symptoms, even if they were already depressed when data collection first started.

Decision Tree Learning is one of the Machine Learning's most used and practical methods for inductive inference, being the most popular among the inductive inference algorithms in large areas such as health, finances, learning to diagnose medical cases, or evaluate possible cases of financial risk. This method aims at approximating functions of discrete values with robustness on data with possibility of noise and learning disjunctive expressions. It is represented by a decision tree with possibility to associate if-then rules to improve human readability.

Decision Tree learning classifies instances from the root of the tree to a leaf node that provides the class instance. Each node of the tree specifies the test of some attribute of the instance. An instance is classified starting at the root node of the tree and testing the attribute related to this node, following the branch that corresponds to the value of the attribute in the instance at matter. This process is then repeated for the subtree below until a leaf node is reached [35].

The Learned Decision Tree shown in figure 5.5 has the objective of classifying the level of risk of heart failure, by checking if it is appropriate common symptoms. The instance $< Systolic\ blood\ pressure \leq 91, Age \leq 62.5, Sinus\ tachycardia\ present = yes >$ would be sorted down to the leftmost branch of this decision tree and then will be classified as a negative instance. The tree predicts that $Risk = High$.

In sum each path from the root of the tree that goes to a leaf corresponds to a conjunction of attributes tests and the tree itself a disjunction of these conjunctions.

A Classification Tree is a classification method which uses training - or historical - data to construct a decision tree. The training set consists of a classified subset of data for a given sample. This learning sample is used to generate a decision tree which is able to classify untrained data. The decision tree divides the trained sample into subsets by asking boolean questions. At each iteration, these questions allow choosing one of the two subsets, according to the answer. That is, asking “Did the elder send over 15 text messages on a given week?” divides the sample data into two: those that did send over 15 text messages, and those that didn’t. The CART algorithm attempts to find the questions that provide the best division of the sample into parts as much homogeneous as possible. This process is repeated until all samples in the given subset are of the same class, which correspond to the tree’s leafs, shown in yellow in Figure 5.5.

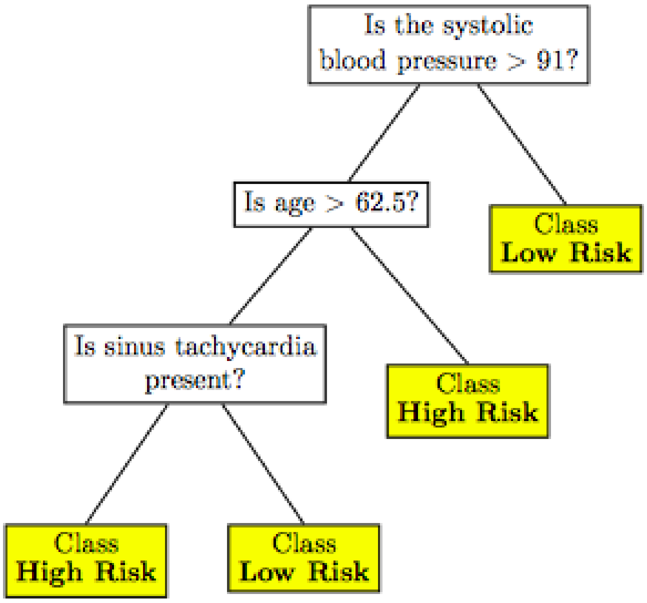


Figure 5.5: An example of a Learned Decision Tree, with leafs shown in yellow (reprinted from [52]).

.

Maximum homogeneity is defined by an impurity function. As such, at each node the CART algorithm solves a maximisation problem. There are several impurity functions, but only two are generally used, the *Gini splitting rule* and *Twoing splitting rule*. Scikit-learn, our framework of choice, uses the *Gini splitting rule*. This rule looks for the largest class in the dataset and tries to isolate it from all the others.

Figure 5.6 shows a decision tree generated from four classes, A, B, C, and D, with sizes 40, 30, 20, and 10, respectively. The Gini splitting rule first separates class A from the rest of the dataset, after that B, and so on. Such homogeneous splits are not always possible, in which case the dataset is split into less optimal groups 5.6.

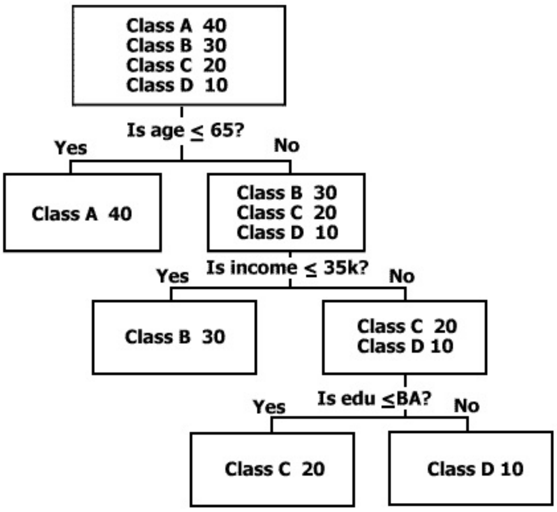


Figure 5.6: Gini rule example(reprinted from [14])

Figure 5.7 shows an example of a tree generated by the application. In this case, the field of activities for the past two weeks is evaluated first. In case the elder spent less than (approximately) 157 minutes practicing some form of activity, then the classification takes the right branch, which is a leaf, and indicates that there are depressive symptoms. Otherwise, the number of not answered calls for the past two weeks is analysed instead. In this case, the tree classifies the elder as having depressive symptoms if there are at least 16 not answered calls in the past two weeks. Next the field of not answered calls is analysed for the past two weeks, but in this case no conclusion is taken in either case, as neither child is a leaf. For the sake of brevity, we'll analyse the right branch, which leads to the nearest leaf indicating the absence of depressive symptoms by analysing, again, the field of not answered calls for the past two weeks. In contrast with the previous analysis, less than 11 calls missed indicate that there are no depressive symptoms.

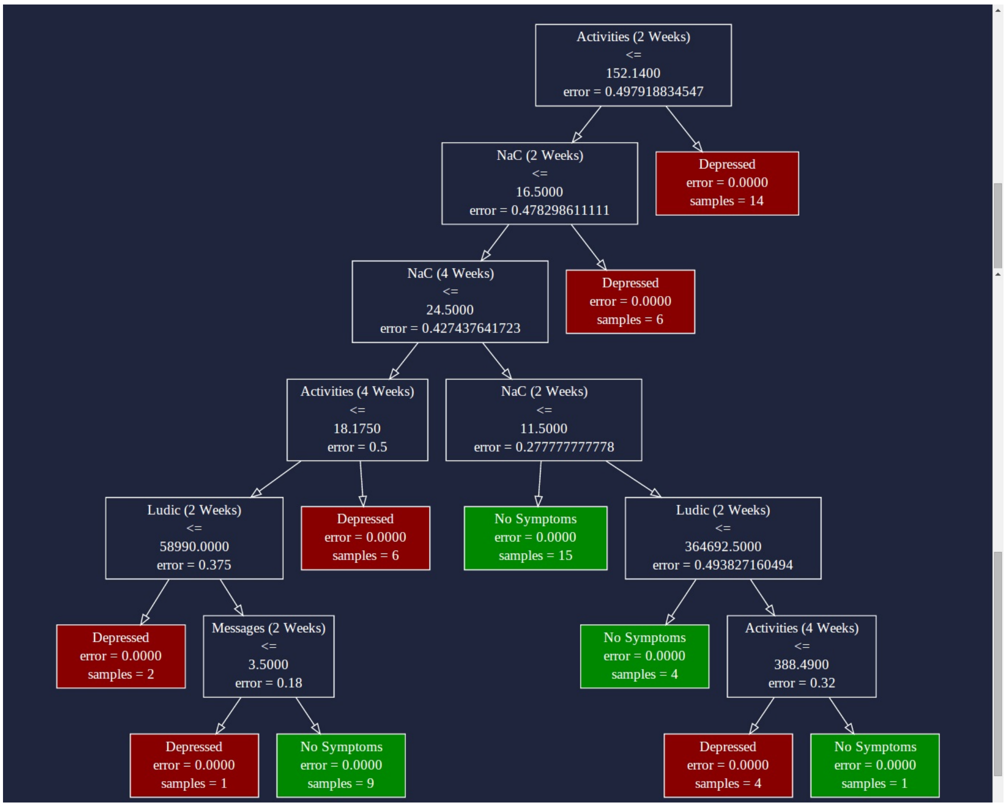


Figure 5.7: Visualization of a Learned Decision Tree.

5.5 Evaluation

The evaluation step was probably one of the greatest impediments we found on this project. Not only because there's no real data, but also because our solution remains to be tested in a real-life situation. Our tests consisted mainly of showing our results to a trained psychologist asking whether the results seemed reasonable, and whether our implementation would suit the standard depression-detection process. The feedback we got on that regard on the psychologist's part was good, even though it was agreed that further testing was still necessary. As such, most of this process was delegated to future work, making our implementation not suitable for release. In order to finish evaluating our results, we would require three things:

- access to the SmartCompanion database;

- further tests under psychologist guidance;
- and real-life situation tests.

Only then could a release version be considered.

5.6 Deployment

The deployment phase is described in detail in chapter 6.

Chapter6

Technical Aspects and Implementation of DepSigns

In this chapter we present a descriptive view of how we structured and implemented this project. We start by exposing and explaining the technologies used and then proceed to correlate them with our own models.

6.1 Used Technologies

Throughout this section, we describe used technologies to better understand the DepSigns implementation.

6.1.1 Scikit Learn

Scikit Learn is an open source machine learning library for the Python programming language. In this library we have the possibility to use several Machine Learning algorithms such as Decision Tree Learning, Support Vector Machines, or Naive Bayes, among others. This library is designed to interoperate with the numerical and scientific Python libraries *NumPy* and *Scipy* [9].

6.1.2 Web Services

The World Wide Web consortium defines a Web service as "a software application identified by an Uniform Resource Identifier (URI), whose interfaces and bindings are capable of being defined, described, and discovered as Extensible Markup Language (XML) artifacts. A Web Service supports direct interactions with other software's agents using XML-based messages exchanged via Internet-based protocols"[1]. However, the RESTful architectural style is also applied to the development of web services [18]. RESTful is an architectural style which imposes some restrictions on the components of the system. Specifically, it is based on a Representational State transfer architecture which consists of a network of resources (commonly called "API"). RESTful services are based on the HTTP protocol and therefore they are stateless. Also, these services use the standard HTTP verbs (GET, PUT, POST, and DELETE) in association with a given resource providing meaningful APIs and convenient communication protocols[31]. We implement a RESTful web service with additional HTML views for data visualization.

6.1.3 Model-View-Controller

The Model-View-Controller (MVC) is a design pattern for implementing applications. This paradigm is divided in three parts [44]:

- *Model* represents an application's current state, by storing and managing a graph of objects or data. The model is responsible for the storage part of an application and, sometimes, an abstraction over that data which allows other layers (such as the Controller) to retrieve data without knowing the specifics of the underlying model.
- *View* views are responsible for the visual part of the application, commonly called *user interface*. Therefore a view must render and manage events related with the rendered elements (such as *clicks*, or text input). When such events occur, the view component is also responsible for delegating to the appropriate section of the Controller. In the specific case of web services, views usually consist of

HTML pages which render a given resource, although they are not limited to it.

- *Controller* is an abstract module which provides indirect model-to-view and view-to-model communication. As such, changes to the view need not reflect on the model, and vice-versa. When an event occurs in either module, the controller is responsible of dispatching the event appropriately. As such, the Controller is a central module in the MVC design, and usually provides the core functionallity of the application.

6.1.4 Database

A Database is an organized collection of data or information for a fast search and retrieval by a computer. This collection of data is structured to facilitate the storage, retrieval, modification and deletion of the information in association with several data-processing operations. A Relational Database Management System (RDBMS) is program that manages such databases. There are various types of RDBMS' and we opted for MySQL as it is the most popular open-source RDBMS. MySQL is based on the Structured Query Language *SQL*, commonly used in web servers. The main goal is to provide access of different types of information contained in the database[8].

6.1.5 Backend

In a web service the backend consists of the server side implementation and therefore it is usually where the core functionallity is implemented. We use *django* to implement our backend. Django is maintained by the *Django Software Foundation* and it is a free and open source web application framework, written in Python, which follows the model–view–controller architectural pattern. This framework allows to use basic modules, classes and tools to quickly develop web applications. It also provides a template framework allowing conversion of HTML and others files into templates. Django also provides an automatically generated administration interface to manage data models, or authentication control, among many other features. Django follows the DRY principle (Don't Repeat Yourself) in which preventing code repetition is a goal. Django is rapidly becoming popular among web developers.

6.1.6 Frontend

The frontend is the part of the application that is closest to the user, usually some sort of user interface. When implementing a web service it is common to use HTML and, CSS, and JavaScript to implement such interfaces.

- **Hypertext Markup Language**(HTML) is a standard markup language to create web pages, which browsers interpret and render. HTML allows images and objects to be integrated in forms and presents a semantic structure[43].
- **Cascading Style Sheets**(CSS) is integrated with HTML and used to alter the look of the HTML elements. CSS allows the separation of document content from presentation, including elements such as fonts, colors and the layout. This style sheet language provide more control and specification of presentation characteristics[24].
- **JavaScript** is a programming language commonly used to control the web browser, alter the document content, or for asynchronous communications. The implementation is made on the client-side with the aim of interacting with the user[10].
- **Asynchronous JavaScript and XML**(AJAX) is a style of web development. It is not a programming language but a concept. The development is made on the client-side by sending requests to the server without the necessity for a *postback* or a page refresh, allowing the web server to communicate dynamically without loading several pages via XML request[53].

6.1.7 Chart Rendering

We use *Highcharts* to render all the data projection. It consists of a chart rendering library developed by *Highsoft AS* that allow us to implement interactive and dynamic charts inside a web application or web site[4]. Highcharts has been written in pure HTML5 and JavaScript, which allows displaying charts natively in a web browser, without using any plugins. Another strength of Highcharts is that charts are created with SVG or VML (for Internet Explorer), which are vector image formats, which means there's no quality loss if the image is displayed bigger formats.

6.2 System Architecture

Being multi-user in nature, this project was a great candidate for a Web Service. For this reason, we used the most common paradigm for implementing this kind of services on the web, knowing that this methodology has been greatly explored and has been vastly used in a variety of different kinds of projects. Also, being web-oriented was a key factor in the decision of the structure used for this application. Nonetheless, the key features are related to how we process the data and achieve depression-detection-related results, and that's where our main focus will be throughout the text.

One of the most typical approaches when creating a Web Service is using the Model-View-Controller paradigm, even though there are other options. Each of these three conceptual modules (Model, View, and Controller) are further divided in functional modules, which constitute the implementation of our application. Each module has its own very specific purpose and is capable of performing its duty independently of the others; meaning that changes to one module will not require changes to others, for as long as the input and output formats remain the same.

Next, we will explain what modules compose this project, what their goals are, and how they relate to each other.

6.2.1 Structure

Figure 6.1 shows an overview of how the project was structured, listing all functional modules and how they relate to each other.

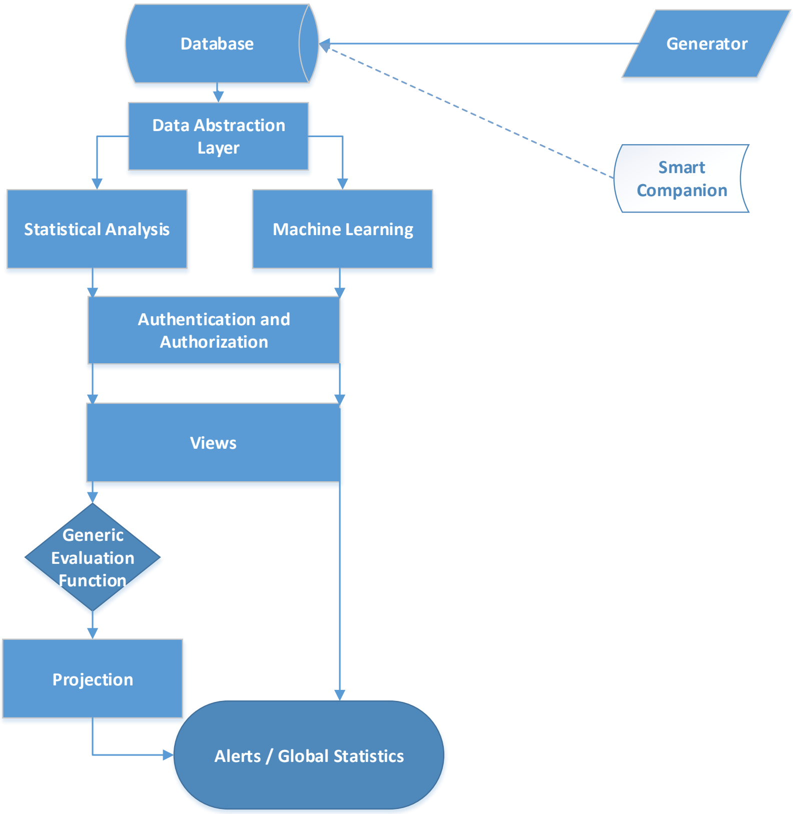


Figure 6.1: DepSigns System Architecture

The overall structure of the project can be divided in three main parts:

- **Model**, which is composed of the Database, the Database Generator, and Data Abstraction Layer. This module is responsible for handling all data retrieval and standardization, making it easy for the Machine Learning and Statistical Analysis to process results.
- **Controller**, which is composed of the modules that handle Machine Learning, Statistical Analysis, Authentication, and Authorization. This module is responsible

for managing the relationship between the Views and the Model, impeding any direct communication between the two. This way we can assure proper results and data consistency.

- **View**, which is composed of all modules responsible for presenting the data to the user. This includes graphic rendering and data layout.

The modules herein described are abstract, that is, not physical components of the system. However, their constituents, as depicted in Figure 6.1, are actual implementations and deserve a deeper approach. For this reason, the following sections describe each of them individually and how they relate to each other.

6.2.2 Database generation

The standard way of creating a database in django[21] consists of listing several models (python classes, which inherit from *django.db.models.Model*) in a single file. These models list the fields which are to be created in each entity and automatically generate a database table (entity). Django then provides a simple way of synchronizing these models with the database:

```
python manage.py syncdb
```

The previous command will search for every model present in the project and generate a database entity for it. In the appendix B, listing B.1, the code shows every model we created for this database. For a description of the purpose each one fulfills see Chapter 5, Section 5.2.

The Database Generator consists of a single unit file which only needs to run once. This script fills the database tables with data, as described in the Section 5.2.

Each table needs its own generator because each one represents a different model. In order to ease the generation process we created a base class, called *Generator*, which already provides common functionality, such as generating dates, times, and time intervals. This class does not relate to any model. Each of the other generators then inherits from this base class and provides specific functionality for the model it represents. We start by explaining the generators for the two central modules in our design: users and moods.

The user generator generates all permutations from two lists of names: one of given names and other of surnames. We also provide a year interval which will limit the minimum and maximum age of each generated individual, although this is quite irrelevant for our study, as all users are considered as being elders. For each permutation we generate a different entry, as we can see in Listing 6.1.

```
class UserGenerator(Generator):

  def generate_names(self):
    for given in self.given_names:
      for sur in self.sur_names:
        yield given + ' ' + sur

  def generate(self):
    l = []
    for name in self.generate_names():
      user = User(name=name, birth=self.generate_date(self.min_year, self.
          max_year))
      l.append(user)
      user.save()
    return l
```

Listing 6.1: The *UserGenerator* class

We then proceed by defining the mood generator. As all the users have already been created, we now generate mood values using the process described in Section 5.2. In review, we create random alpha and beta parameters, between 1 and 5, for the beta distribution function, which generates values in the interval $[0;1]$, with a given distribution. Moods are then discretized into four categories: *Bad*, *Not Well*, *Fine*, and *Very Well*. Also, the *start* and *end* parameters indicate the period of data that will be generated. Between these two dates, we will generate an entry for each day. This also applies for every other generator from here on.

```
class MoodGenerator(Generator):
  def generate(self):
    for user in self.users:
      user.alpha = random() * 5 + 1
      user.beta = random() * 5 + 1
      start, l = self.start, []

      while start != user.end:
        mood = betavariate(user.alpha, user.beta)

        if mood < .25: mood = 'B'
        elif mood < .50: mood = 'N'
        elif mood < .75: mood = 'F'
        else: mood = 'V'

        mood = Mood(user=user, day=start, moodvalue=mood)
        mood.save()
        l.append(mood)
        start += timedelta(days=1)

      user.moods = l
```

Listing 6.2: The *MoodGenerator* class

The following code defines the generator for activities. Because every remaining entity generator uses pretty much the same method, we define the *ActivityGenerator* as a base class that implements the *make(self, user, start, end)* method (cf. Listing 6.3,line 31). This is overriden by subclasses to change the model instance that is created in the process. So, to generate the remaining entities, we use the same α and β parameters for the beta distribution function, as randomly generated when creating the moods. The moods are then, again, used as a threshold for obligation compliance, as previous described (See Chapter 5, Section 5.2).

```
1  class ActivityGenerator(Generator):
2
3    def generate(self):
4      self.count_ok = { 0: 0, 1: 0, 2: 0 }
5      self.count_not = { 0: 0, 1: 0, 2: 0 }
6      for user in self.users:
7        start, index = self.start, 0
8        while start != self.end:
9          mood, index = user.moods[index].moodvalue, index + 1
10         for i in range(2):
11           prob = betavariate(user.alpha, user.beta)
```

```
          self.generate_period(i, prob, user, start, mood)
        prob = 1 - betavariate(user.alpha, user.beta)
        self.generate_period(2, prob, user, start, mood)
        start += timedelta(days=1)

  def generate_period(self, period, prob, user, day, mood):
    if mood == ’B’: threshold = .90
    elif mood == ’N’: threshold = .70
    elif mood == ’F’: threshold = .50
    else: threshold = .25
    if prob < threshold:
      return
        start_hour, end_hour = self.time_according_to_period(period)
    start, end = self.generate_time_2(start_hour, end_hour)
    day = ("%s" % day).split(’ ’)[0]
    start, end = "%s %s" % (day, start), "%s %s" % (day, end)
    act = self.make(user, start, end)
    act.save()

  def make(self, user, start, end):
    return Activity(user=user, start=start, end=end)
```

Listing 6.3: The *ActivityGenerator* class

All other generators simply override the *make* method to change the returned instance. The exception goes to the *LudicActivityGenerator* which also generates a score value.

```
class LudicActivityGenerator(ActivityGenerator):
  def make(self, user, start, end):
    prob = betavariate(user.alpha, user.beta) * 100
    return LudicActivity(user=user, start=start, end=end, score=prob)

class LocationGenerator(ActivityGenerator):
  def make(self, user, start, end):
    return Location(user=user, check_in=start, check_out=end)

class NotAnsweredCallGenerator(ActivityGenerator):
  def make(self, user, start, end):
    return NotAnsweredCall(user=user, when=start)

  def generate_period(self, period, prob, user, day, mood):
    return super(NotAnsweredCallGenerator, self).generate_period(period, 1 - prob, user, day, mood)

class CallGenerator(ActivityGenerator):
  def make(self, user, start, end):
    return Call(user=user, start=start, end=end)

class MessageGenerator(ActivityGenerator):
  def make(self, user, start, end):
    return Message(user=user, when=start)
```

Listing 6.4: The remaining generator classes

The raw data is not in a convenient format for use in the application. On top of that, it was a well known fact from the start of this project that the data source could change in the future (that is, use of the Smart Companion database). As such, we created a Data Abstraction layer over the Database layer which allows to both abstract from the data source and convert the data into a more suitable format.

6.2.3 Authentication and Authorization

Being multi-user in nature, this application required an authentication and authorization process.

Authentication consists in identifying someone as actually being whom he/she claims to be. Each user is required to have a username and password. We use django as an underlying mechanism to implement this feature, as it already provides the necessary routines to do so.

Authorization, on the other hand, consists in assigning each authenticated user a set of permissions. Our service only implements two kinds of users with different levels of permissions : psychologists and caregivers.

The psychologist is allowed to:

- access the main page, where main statistics, general alerts, a calendar, among others are present;

- search patients by name;
- provide manual feedback to build the training dataset given to the decision tree learning algorithm;
- see the learned decision tree, as generated by the CART algorithm;
- see individual, personal, alerts for each elder;
- see data history;
- choose the centrality tendency metric used for each patient.

The caregiver has a reduced set of privileges:

- access to the personal page of only one elder;
- see field-wise and population-infered alerts generated for that elder.

By reducing the amount of information we allow the caregiver to see, we protect the seniors personal information. This data, however, is fundamental for the psychologist to derive conclusions on the seniors mental state.

6.2.4 Views

This web service is composed of a set of pages which allow its users to explore the implemented functionality. Next we describe these pages, their URL mappings, authorization requirements, and objectives.

The set of URL mappings constitute the service's API. Each URL serves a different purpose, as described in table 6.1.

URL	Purpose
/	The main page
/login	User authentication
/logout	User logout
/register	User registration
/feedback	Allows providing learning feedback, use to construct the learning decision tree
/patient	See patient history and personal alerts
/analysis	See a renderization of the learned decision tree, generated from psychologist feedback
/patient_list	See a list of available patients, as well as searching patients by name

Table 6.1: Purpose of each URL mapping

Some of these URLs return HTML, which consists of a visualization of some sort, others use AJAX to perform some distinct operation. We implement this by using django Views, which consists of classes that allow specifying how each URL behaves and what HTTP methods it supports. We now describe each view individually.

6.2.4.1 Main Page

The main page has tree different views, depending if the user is:

- not authenticated;
- not registered;
- authenticated.

In case the user is not authenticated the main page shows a login form. The login form allows the user to identify him/herself.The figure 6.2 shows this page.

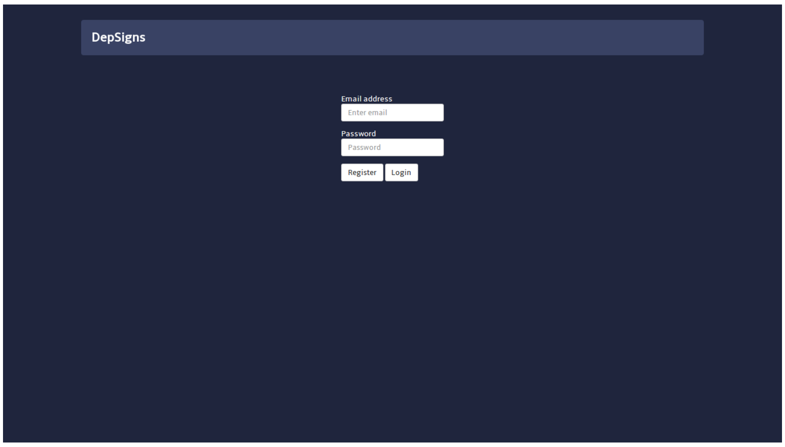


Figure 6.2: Login page

If the user is not registered yet this page also offers a registration form, as shown in figure 6.3.

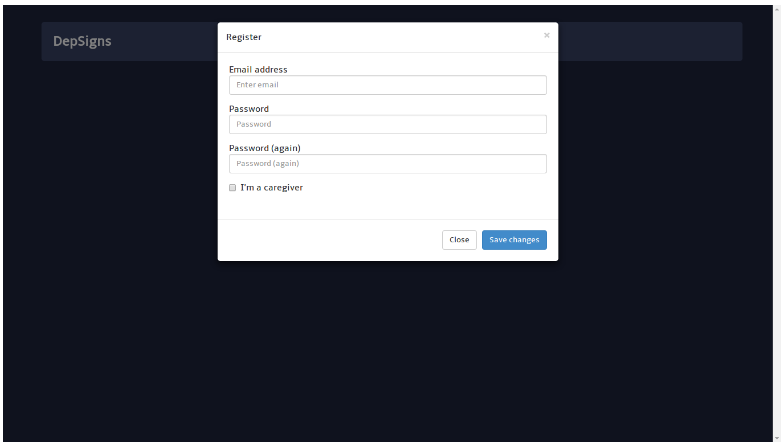
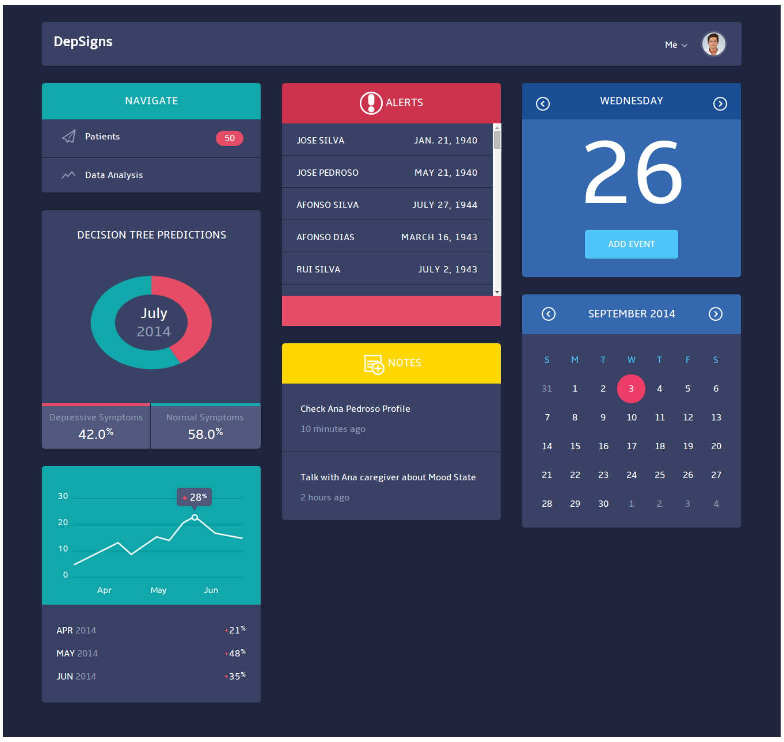


Figure 6.3: Registration form

If the user is already authenticated, the behavior for this page varies according to the type of user. In case the user is a caregiver, he/she will be redirected to the page describing the patient that has been assigned to him/her. If the user is registered as a psychologist the page will show a navigation menu, which allows navigating to other pages, a list of alerts identified by the Decision Tree, the set of predictions made by the Decision Tree, and other sections which demonstrate features to possibly implement in future versions of the service. Figure 6.4 shows a screenshot for the psychologist's main page, where several features can be identified. The navigation menu (top left) has links for the list of patients and the learned decision tree projection; the lists of alerts (top center, in red) gives an overview of patients with infered alerts; Decision

Tree Predictions (the section with the same title) shows how the patients are classified with (*Depressive Symptoms* and *No Symptoms*); all other sections are future work and do not provide any functionality whatesoever.



JOSE SILVA	JAN. 21, 1940
JOSE PEDROSO	MAY 21, 1940
AFONSO SILVA	JULY 27, 1944
AFONSO DIAS	MARCH 16, 1943
RUI SILVA	JULY 2, 1943

Depressive Symptoms	Normal Symptoms
42.0%	58.0%

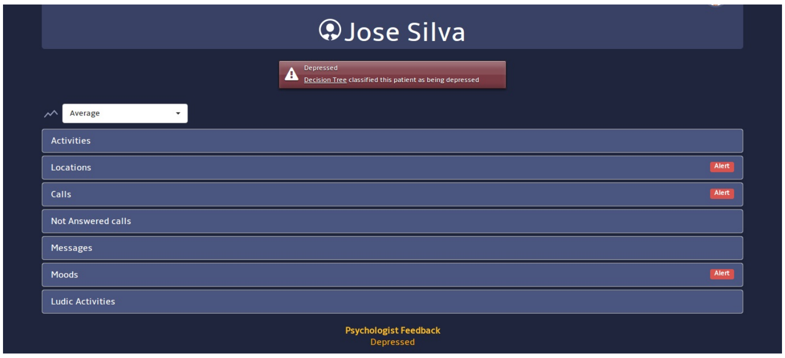
APR 2014	+21%
MAY 2014	+48%
JUN 2014	+35%

Check Ana Pedroso Profile
10 minutes ago

Talk with Ana caregiver about Mood State
2 hours ago

S	M	T	W	T	F	S
31	1	2	3	4	5	6
7	8	9	10	11	12	13
14	15	16	17	18	19	20
21	22	23	24	25	26	27
28	29	30	1	2	3	4

Figure 6.4: Psychologist's main page.



Depressed
Decision Tree classified this patient as being depressed

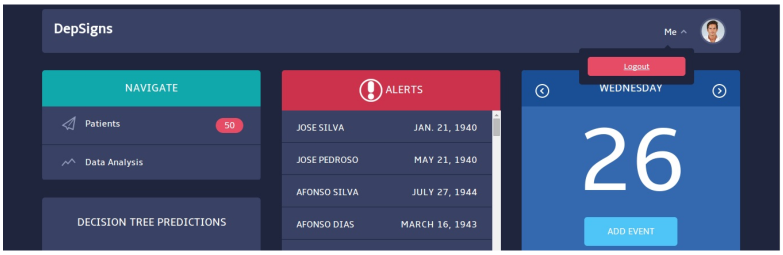
Activities
Locations
Calls
Not Answered calls
Messages
Moods
Ludic Activities

Psychologist Feedback
Depressed

Figure 6.5: Caregiver's main page.

6.2.4.2 Logout

The logout view causes the user to be unauthenticated and redirects the page to the main page.



JOSE SILVA	JAN. 21, 1940
JOSE PEDROSO	MAY 21, 1940
AFONSO SILVA	JULY 27, 1944
AFONSO DIAS	MARCH 16, 1943

Figure 6.6: Logout option.

6.2.4.3 Feedback

The feedback view uses AJAX to allow the psychologist to provide learning feedback to the service. The psychologist indicates whether, in his/her opinion, the patient belongs to the class that shows depressive symptoms or not. This feedback is then used to construct the training dataset provided to the decision tree learning algorithm (CART). No visualisation whatsoever exists for this URL mapping, being an AJAX call. This view can be called from the page showing patient history and personal alerts.

This view also renders the visualization of the learned decision tree. By rendering the tree everytime feedback is provided, we are rendering it every time there are changes. As such, we need not render the same image several times, by caching this renderization on the server. This process is described in Chapter 5, Section 5.4.

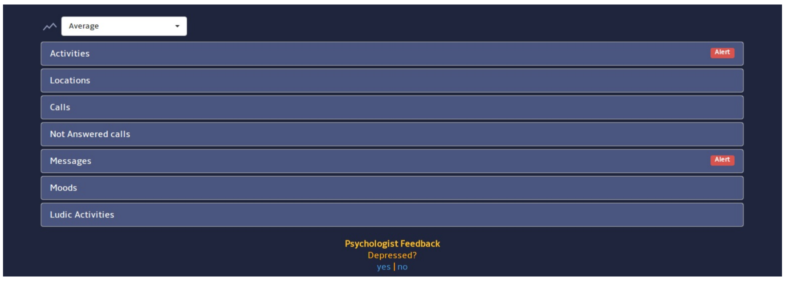
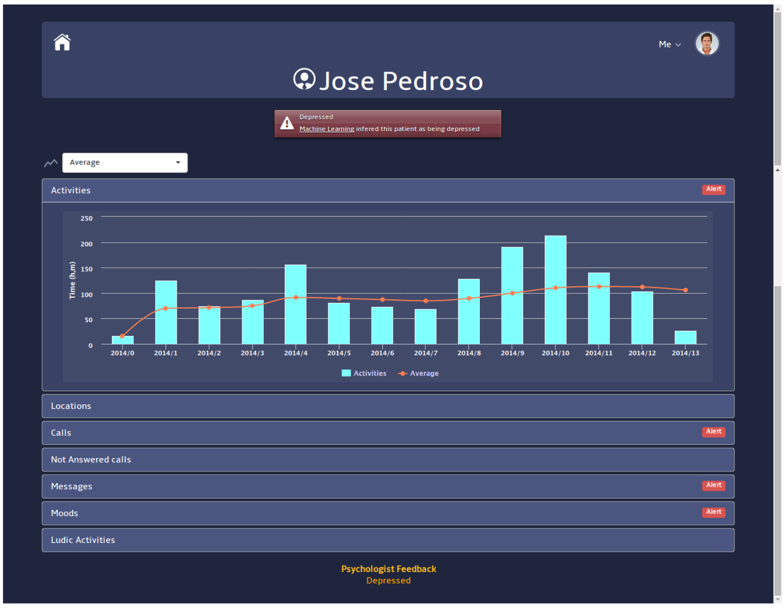


Figure 6.7: Feedback psychologist option.

6.2.4.4 Patient

In this page we show a detailed history of the elder's activities. Each type of registered activity may also issue an alert, indicating that the statistical analysis found deviances regarding the centrality metric for the given field. Both the psychologist and caregiver can see these alerts, but only the psychologist can see the details for the seniors recent history. Also, this page allows choosing the generic evaluation function, which might change the field-wise alerts shown. Finally, two sections exist dedicated to the Machine Learning process. At the top of the page there's an indication of the class inferred by the learned decision tree (that is, either the elder shows depressive symptoms or none at all). At the bottom of the page, there's a section dedicated to psychologist feedback. Here, if the psychologist has not yet given feedback about this elder, he/she can do so. If the psychologist has already provided feedback, their conclusions are shown here, indicating what the feedback was at the time it was given. Note, however, that these two processes are independent from each other. The psychologist needs not to provide feedback for all elders in order to see Decision Tree inference for any given senior. Still, the psychologist can provide feedback to be later incorporated into the training data supplied to the decision tree learning algorithm. Figure 6.8 depicts the patient page, showing history for the seniors Activities, and alerts for each individual field. At the top left corner there's a combo box that allows the psychologist to choose the evaluation function. At the center top there's an indication of the Decision Tree classification; red indicates depressive symptoms and green the absence of symptoms. At the bottom there's an indication that the psychologist has already indicated this patient as being depressed



Jose Pedroso

Depressed
Machine Learning infered this patient as being depressed

Average

Psychologist Feedback
Depressed

Figure 6.8: The patient page.

6.2.4.5 Analysis

This view allows the visualization of a renderization of the learned decision tree. The process of rendering is performed when the psychologist provides learning feedback, so this view merely returns the rendered tree. Additionally, we process the tree to identify which nodes indicate depressive symptoms and which indicate the absence of symptoms. Also, the tree is simplified to provide a better overview of the flow of execution to the psychologist. This process is described in Chapter 5, Section 5.4. The renderization is performed using SVG, and there's a PDF download available. Figure 6.9 shows the final result.

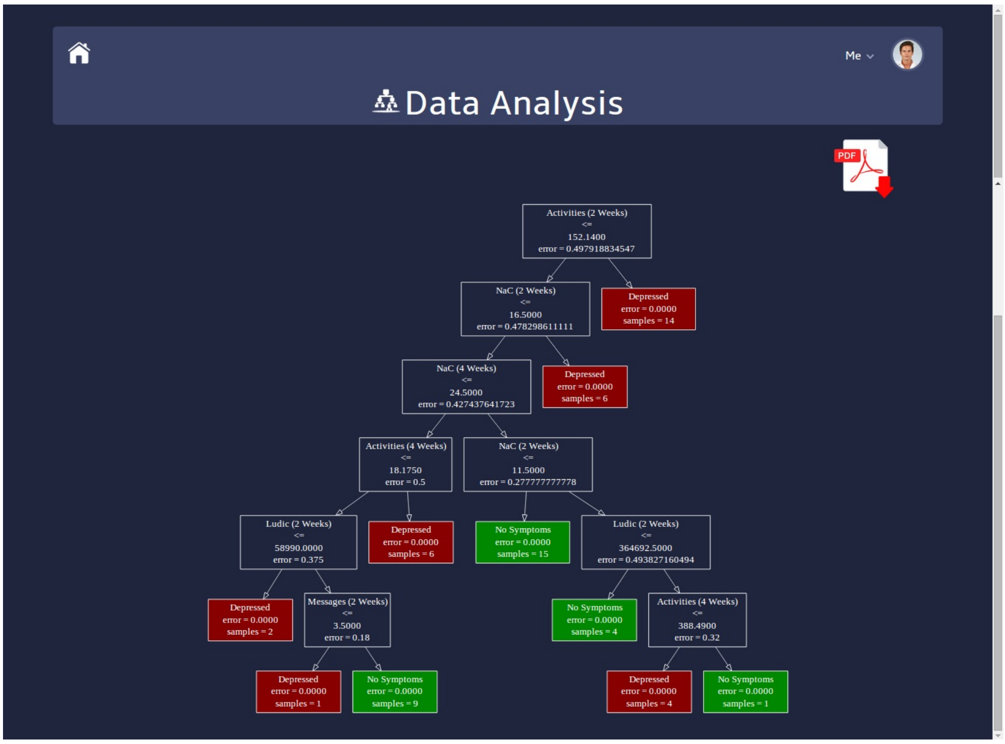
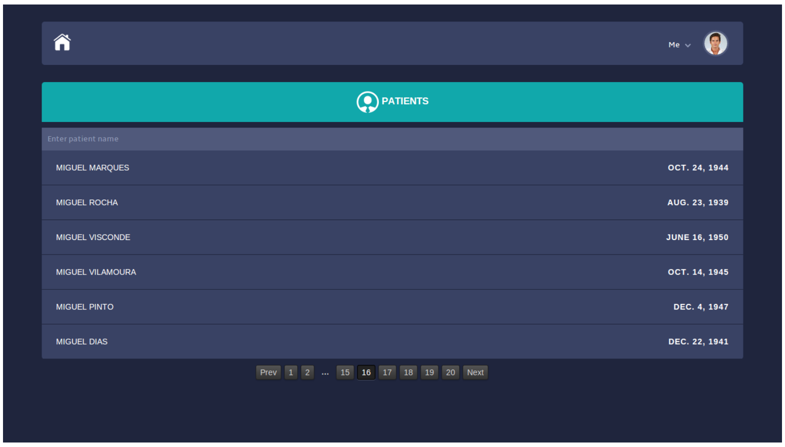


Figure 6.9: Visualization of the Learned Decision Tree.

6.2.4.6 Patient List

The patient list allows the psychologist to see a list of all registered patients, as well as search them by name. Figure 6.10 shows this.



MIGUEL MARQUES	OCT. 24, 1944
MIGUEL ROCHA	AUG. 23, 1939
MIGUEL VISCONDE	JUNE 16, 1950
MIGUEL VILAMOURA	OCT. 14, 1945
MIGUEL PINTO	DEC. 4, 1947
MIGUEL DIAS	DEC. 22, 1941

Figure 6.10: Page with the list of patients, and a search bar.

6.2.5 Projection

The implemented visualization tool uses the highcharts framework to render the charts. Although the generated data format is common to both the Statistical Analysis and Machine Learning Analysis, it does not share common grounds with the highcharts framework requirements. For this reason, we first need to convert the data to a format that highcharts recognizes. This is done on the client side, by the function depicted in 6.5.

```
function normalize_graph_data(data, context)
{
  var result = [];
  var r_data = { 'name': context, 'data': [], 'type': 'column','color':'#80FFFF' };
  var r_spline = { 'name': get_evaluation_function_name(), 'data': [], 'type': 'spline' };
  var spline_values = [];

  for (var it in data) {
    var value = data[it]['weight'];
    var rounded_value = round2(value);
    var spline;

    spline_values.push(value);
    spline = compute_spline_f(spline_values);
    spline = round2(spline);

    r_data['data'].push(rounded_value);
    r_spline['data'].push(spline);
  }

  return [r_data, r_spline];
}
```

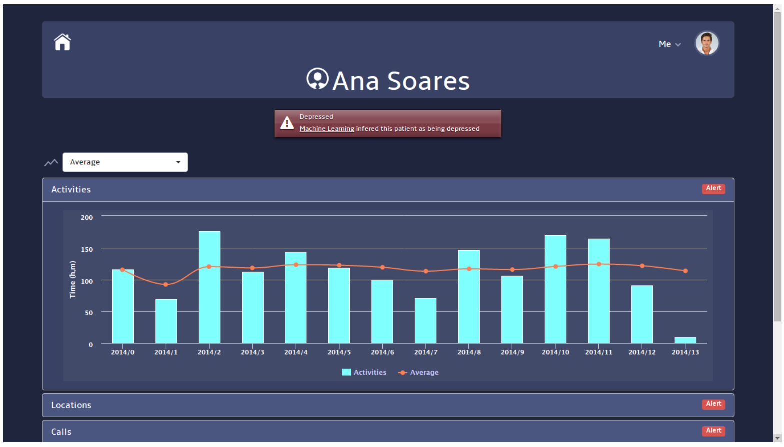
Listing 6.5: Normalize chart data function

The data returned by this function consists of an array containing data and spline values, which can now be used to render the charts. This array is then passed to the highcharts API in the call to the highcharts jQuery plugin, using the *series* argument which is the standard parameter for passing renderable data. After the data is normalized, the highcharts framework takes over and the rest of the process is automatic.

6.2.6 Alerts

There are two kinds of alerts which can be generated: field-wise and field-agnostic. Field-wise alerts are generated by the Statistical Analysis, providing an insight of how the elder matches his/hers usual habits. Field-agnostic alerts, on the other hand, are generated by the Decision Tree and provide a means of comparing the individual's physical and mental state by comparison with the general population. These alerts are shown to both the psychologist and the caregiver.

The field-wise alerts are generated by the method described in 6.6. Here, we compare the elder's data with the splines generated previously (see Chapter 5, Subsubsection 5.2.4), starting from the end, and counting *N* weeks back, which, by definition, consists of 2. Values below the spline line contribute to generate an alert, unless the spline is considered to be inverted, which indicates that values below the spline contribute positively to the elder's state of mind; this only happens for the field of *Not Answered Calls*, in which we consider that the fewer calls are missed, the better. If the count reaches two, then an alert is shown for the given field, as shown in Figure 6.11.



Depressed
Machine Learning infered this patient as being depressed

Figure 6.11: Example of field-wise alert

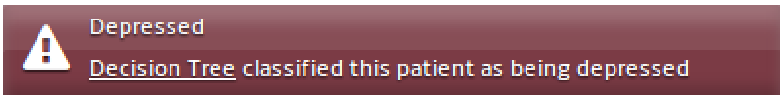
```
function check_alerts(header, data, spline_inverted)
{
  var N = 2;  // Number of weeks to consider
  var count = 0;
  var spline = data[1].data;
  data = data[0].data;

  for (var it=data.length - 1 ; it >= 0 && it >= data.length - N ; it--) {
    if (!spline_inverted && data[it] < spline[it] || spline_inverted && data[it] > spline[it]) {
      count++;
    }
    else break;
  }

  if (count >= N) {
    show_alert(header);
  }
}
```

Listing 6.6: Check alert function

Field-agnostic alerts, on the other hand, are originated from the server and need not be processed any further then described in Chapter 5, Section 5.4. A boolean field is returned from the server indicating whether the alert should or should not be shown. Because we do not need the javascript runtime to generate this alerts, we used django's template system to check for the boolean flag. Listing 6.7 shows this. If the elder has been classified by the Decision Tree as having depressive symptoms, figure 6.12 is shown, otherwise figure 6.13 is shown in its place.

```
{% if inference %}
<div class="nn-alert-container">
  <div class="nn-alert depressed">
  <i class="fa fa-exclamation-triangle"></i>
  <strong>Depressed</strong>
  <p><u>Machine Learning</u> infered this pacient as being depressed</p>
  </div>
</div>
{% else %}
<div class="nn-alert-container">
  <div class="nn-alert not-depressed">
  <i class="fa fa-check"></i>
  <strong>No Depressive Symptoms</strong>
  <p><u>Machine Learning</u> infered this pacient as not having depression symptoms</p>
  </div>
</div>
{% endif %}
```

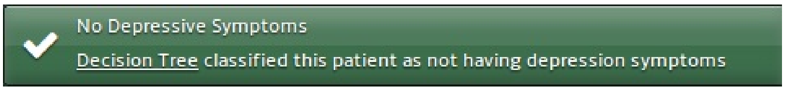
Listing 6.7: Inference code



Depressed

Decision Tree classified this patient as being depressed

Figure 6.12: Decision Tree classified as depressive.



No Depressive Symptoms

Decision Tree classified this patient as not having depression symptoms

Figure 6.13: Decision Tree classified as no depressive.

6.2.7 Decision Tree Predictions

One of the main purposes of this work was to use a Machine Learning algorithm to predict depressive symptoms. As such, we have classified part of the population as depressed and not depressed building the training data supplied to the Decision Tree learning algorithm. Afterwards, run the learned Decision Tree over all the population and obtain the class predictions for all the individuals. This process allows us to know how many individuals are showing depressive symptoms, and how many are not, including the training samples. We believe this can be useful at several levels. On one hand, the psychologist is able to analyse the global status of the elders under his/hers care. On the other hand, it could allow psychologists to compare this data and, hopefully, discuss and compare the methodologies being used, comparing the results with each other. We believe this could help in identifying better treatments by comparison.

We implement this by iterating the list of registered individuals (users) and using the *infer_user_symptoms* method (described in Chapter 5, Section 5.4) on each; this is enough, as all elders that fail to pass this test are considered to have no depression symptoms. Therefore, we list all elders having depressive symptoms, as shown in listing 6.8. Given this list, it becomes trivial to count the number of patients lying on each class, given the total number of patients. The server then returns the percentage of depressive, and non-depressive, over the total number of individuals. We then use the Chart.js framework to render the doughnut graph shown in figure 6.14, as described in listing 6.9, where *count_depressed* indicates the number of patients with depressive symptoms and *count_not_depressed* indicates the number of patients without.

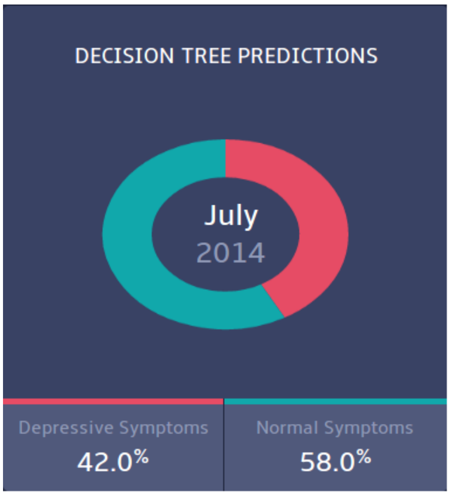


Figure 6.14: Decision Tree Predictions chart.

```python
def infer_depression_list():
  return [usr for usr in User.objects.all() if infer_user_symptoms(usr)]
```

Listing 6.8: Code for inference depression list

```javascript
var doughnutData = [
  {
    value: {{ count_depressed }},
    color:"#E64C65"
  },
  {
    value : {{ count_not_depressed }},
    color : "#11A8AB"
  },
];
var myDoughnut = new Chart(document.getElementById("canvas").getContext("2d")
    ).Doughnut(doughnutData);
```

Listing 6.9: Render Doughnut graphic

Chapter7

Conclusion

We finish by giving an overview of the conducted work in this thesis. We present our main contributions and challenges and discuss further work possibilities.

We decided to incorporate in Depsigns data related to: activity level categorized by time of day (morning, afternoon, and night); locations (how often the user leaves the house); mood (a self-assessment made by the user); communications (calls and messages) and ludic activities (namely, cognitive exercises). We conduct a statistical analysis on the collected data and use it to infer depression symptoms from personal habits. The goal of this statistical analysis is to find deviations from the senior's usual behaviour, by detecting standard patterns in the senior's daily activities, and causing field-wise alerts to be shown to the psychologist or caregiver. To detect the mentioned outliers/deviances, we use an abstract evaluation function which can be configured by the psychologist using this software. The results achieved by conducting this study vary a lot according to the senior's initial condition. The system will not generate an alert for senior's that already showed depression signs when we first started sampling data, since a depressed state will correspond to their normal condition. This is why we also conduct a study using a Machine Learning algorithm. Decision Tree Learning is one of the most used and practical methods for inductive inference, being the most popular among the inductive inference algorithms in large areas such as health or financial, learning to diagnose medical cases or evaluate possible cases of financial risk [35]. We use this method to classify a set of symptoms as depressive/non-depressive condiction. For this purpose, we asked a psychologist to provide feedback on some seniors, indicating whether symptoms existed. We commit this feedback and generate a dataset by correlating the senior's history with the provided feedback. A decision tree is trained using this dataset and the learned model is used to predict senior's conditions on a new dataset. Contrary to the statistical analysis, the decision tree classifies symptoms according to global population standards[45].

7.1 Summary

As it should be clear by now, the psychologist's feedback was essential in this project. Most of our choices towards depressive symptoms detection were based on this guidance - no other study towards that goal was ever conducted. The choice of algorithms therefore still lacks some justification and further analysis. As of this moment, our conclusions were driven from extreme cases, which the psychologist analysed superficially and indicated whether there were depressive symptoms or not, making it clear to us as to whether our inference process was correct. Even though we achieved a 100% success rate, we still lack the confidence to vouch for these results as it is clear to us that some considerings remain unmade, as we will see in Section 7.3. As such, we know not the actual success rate nor is it within our reach to find out, as some basic requirements would have to be met before:

- real-life data;
- tests performed by psychology experts;
- fine tunning, as described in Section 7.3.

We can conclude, however, that our process can at least identify extreme cases, when considering the generated data.

Statistical analysis failed at identifying depressive symptoms when the elder is already admitted in a depressive state, while the Machine Learning process is not suitable for inferring conclusions from personal habits. Even so, psychologist feedback proved great as to how our process identifies depressive symptoms, just as much as it did with the quality of the result.

We consider that we have fulfilled the proposed objectives, perhaps with the exception of the evaluation phase, which lacks work towards validating our results, and other issues discusses next, in Section 7.3.

7.2 Challenges

Gathering data and unprecedent related work were probably the greatest challenges we faced developing this project. Being personal in nature, the collected data is sensitive and imposes privacy issues. On top of that, the current database for the Smart Companion project did not fulfil the necessary requirements, having forced the fabrication of artificial data. This creates difficulties in achieving real-life representations of elder's state of minds, if such is possible at all. As such, psychologist guidance and advice was essential and consisted of an important backbone for the development process, although multidisciplinary projects often impose other kinds of difficulties, such as achieving an understanding when the people involved stem from completely different backgrounds.

Not having found any precedent work of the kind also imposed a great challenge. The choice of algorithms and methodology is not known to be optimal, having pushed much of the development process for future work. Once again, these choices were made under psychologist guidance, but such is not indicative of optimality, and we rather chose an approach which we consider to be *good enough*. We expect future work to be based upon our own, when results will possibly already be available.

7.3 Future Work

It has been stated many times throughout this thesis that nothing can replace real data, and that our simulator is far from being validated. In order for this project to be continued this would have to be one of the first issues to be resolved. Access to the SmartCompanion database would be, therefore, a great contribute to further developing this project. This form of automatic data gathering would be the next step towards making this a complete solution.

However, other issues exist which still need further studying. A Fine tuning the parameters used to generate the tree is also something we still lack attention in. Decision trees are known to produce less-than-optimal results when upper and lower bound parameters are incorrectly setup or, as with in our case, nonexistent. That is, in order to ensure optimality, we would have to limit the tree's depth to both minimum and maximum levels. However, these parameters are not known to us and only by solving the real-life data issue and further studying our results would it be possible to correctly identify and tune such parameters. On top of that, we cannot dismiss the possibility of using other Machine Learning methods. Our method should be tested, by comparing it with other Machine Learning algorithms.

It is also not known which central tendency measure would produce the best results when conducting the statistical analysis. Instead, we allow the psychologist to choose such measure from a set of available functions, which can be further expanded. As such, we propose as future work that different types of measures are put to a test, even by different psychologists, and that the results are analysed and compared. Not only could this process identify the best suited method but also new central tendency measures which could further improve our solution.

We do not limit our future work proposal to what we think our own work lacks attention in, but also to new features. We believe that this should be a complete tool, which a psychologist could use in the day-to-day work life. As such, we also propose the following:

- prioritized alerts, in which the most critical situations would be shown first;
- a calendar, in which the psychologist could associate events with elders, such as scheduling meetings, or take important notes;
- better statistics of the general population. That is, a deeper statistical analysis of the underlying population, which contrasts with the current solution which only provides the percentage of individuals fitting in each class;
- conduct an inquiry and further study other features which would be useful to both

the psychologist and caregiver.

AppendixA

Acronyms

AIDS	Acquired Immunodeficiency Syndrome
DMS-IV	Diagnostic and Statistical Manual of Mental Disorders
GDS	Geriatric Depression Scale
CES-D	Center for Epidemiologic Studies - Depression
HAM-D	Hamilton Scale - Depression
PHQ-9	Patient Health Questionnaire
KD	Knowledge Discovery
KDDM	Knowledge Discovery and Data Mining
KDD	Knowledge Discovery in Databases
DM	Data Mining
DepSigns	Depression Signs Detection through Smartphone Usage Data Analysis
URI	Uniform Resource Identifier
SOAP	Simple Object Access Protocol
WSDL	Web Service Definition Language
HTTP	Hypertext Transfer Protoco
URL	Uniform Resource Locator

API Application Programming Language

HTML HyperText Markup Language

CSS Cascading Style Sheets

AJAX Asynchronous JavaScript and XML

XML eXtensible Markup Language

HTTP Hypertext Transfer Protocol

CRISP-DM CRoss-Industry Standard Process for Data Mining

AppendixB

Source code

```
class User(models.Model):
  name = models.CharField(max_length=100, null=False)
  birth = models.DateField(null=False)
class Activity(models.Model):
  user = models.ForeignKey(User)
  start = models.DateTimeField(null=False)
  end = models.DateTimeField(null=False)

class LudicActivity(models.Model):
  user =  models.ForeignKey(User)
  start = models.DateTimeField(null=False)
  end = models.DateTimeField(null=False)
  score = models.IntegerField(null=False)

class Mood(models.Model):
  MOOD_CHOICES = (
    (’B’,’Bad’),
    (’N’,’Not well’),
    (’F’,’Fine’),
    (’V’,’Very good’),
  )
  user = models.ForeignKey(User)
  day = models.DateField(null=False)
  moodvalue = models.CharField(max_length=1,choices=MOOD_CHOICES)

class Location(models.Model):
  user = models.ForeignKey(User)
  check_in = models.DateTimeField(null=False)
  check_out = models.DateTimeField(null=False)

class NotAnsweredCall(models.Model):
  user = models.ForeignKey(User)
  when = models.DateTimeField(null=False)

class Call(models.Model):
  user = models.ForeignKey(User)
  start = models.DateTimeField(null=False)
  end = models.DateTimeField(null=False)

class Message(models.Model):
  user = models.ForeignKey(User)
  when = models.DateTimeField(null=False)

class DepressionInference(models.Model):
  user = models.ForeignKey(User)
  depressed = models.BooleanField(null=False)
```

Listing B.1: Django models used to create the database

References

[1] Gustavo Alonso et al. Web services. http://link.springer.com/chapter/10.1007/978-3-662-10876-5_5, (Last accessed: August 25, 2014).

[2] Douglas Altman and J. Bland. Standard deviations and standard errors. www.ncbi.nlm.nih.gov/pmc/articles/PMC1255808, (Last accessed: June 10, 2014).

[3] answers.com. Beta distribution. http://www.answers.com/topic/beta-distribution, (Last accessed: March 3, 2014).

[4] HighSoft AS. Highcharts. http://www.highcharts.com/docs, (Last accessed: July 14, 2014).

[5] American Psychiatric Association. Manual de diagnóstico e estatística das perturbações mentais, texto revisto. chapter Perturbações do Humor, pages 345–375. Climepsi Editores, 2006.

[6] American Psychological Association. Older adults' health and age-related changes, 2014. http://www.apa.org/pi/aging/resources/guides/older.aspx, (Last accessed: May 8, 2014).

[7] David Baldwind and Jon Birtwistle. An Atlas of DEPRESSION. Parthenon Pub. Group, New York, 2002.

[8] Oracle Corporation. Mysql. http://www.mysql.com/about/, (Last accessed: August 1, 2014).

[9] David Cournapeau. Scikit learn: Machine learning in python. http://scikit-learn.org/stable/, (Last accessed: June 10, 2014).

[10] Douglas Crockford. Javascript: The good parts. http://www.google.pt/books?id=

PXa2bby0oQ0C&printsec=frontcover&hl=pt-PT&source=gbs_ge_summary_r&cad=0#v=onepage&q&f=false, (Last accessed: August 3, 2014).

[11] Portal da Saúde. Depressão, 2014. http://www.portaldasaude.pt/portal/conteudos/enciclopedia+da+saude/ministeriosaude/saude+mental/depressao.htm, (Last accessed: March 1, 2014).

[12] Instituto Nacional de Estatística. Projeções de população residente em portugal 2008-2060. 2009.

[13] John Devcic. Weighted moving averages: The basics. www.investopedia.com/articles/technical/060401.asp, (Last accessed: July 8, 2014).

[14] Salford Dystems. Do splitting rules really matter? https://www.salford-systems.com/resources/whitepapers/114-do-splitting-rules-really-matter, (Last accessed: August 25, 2014).

[15] ExcelatLife. Learn to control the stress that contributes to depression, 2014. http://www.excelatlife.com/depression.htm, (Last accessed: April 16, 2014).

[16] Fayyad et al. From data mining to knowledge discovery in databases. The Fourteenth National Conference on Artificial Intelligence, pages 34–57, 1997.

[17] Rocío Fernández-Ballesteros. Aging and quality of life, 2008. http://cirrie.buffalo.edu/encyclopedia/en/article/296/(Last accessed: September 1, 2014).

[18] Roy Fielding and Richard Taylor. Principle design of the modern web architecture. https://www.ics.uci.edu/~fielding/pubs/webarch_icse2000.pdf, (Last accessed: August 25, 2014).

[19] António Fonseca. O Envelhecimento – Uma abordagem psicológica. Universidade Católica Editora, 2004.

[20] António Fonseca and Constança Paúl. Envelhecer em Portugal. Psicologia, Saúde e Prestação de Cuidados. Climepsi Editores, 2005.

[21] Django Software Foundation. django. https://www.djangoproject.com/, (Last accessed: July 17, 2014).

[22] Python Software Foundation. pickle - object serialization. https://docs.python.org/2/library/pickle.html, (Last accessed: July 13, 2014).

[23] Healthy Generation. Depressão, 2014. http://www.hgeneration.pt/hnews_201011.php, (Last accessed: April 8, 2014).

[24] Pierre Genevès et al. On the analysis of cascading style sheets. http://dl.acm.org/citation.cfm?doid=2187836.2187946, (Last accessed: August 3, 2014).

[25] Michael Irwin et al. Screening for depression in the older adult, criterion validity of the 10-item center for epidemiological studies depression scale. American Medical Association, 15:1701–1704, 1999.

[26] Fernandes H. J. Solidão em idosos do meio rural do concelho de bragança. Master's thesis, Faculdade de Psicologia e Ciências da Comunicação da Universidade de Lisboa, 2007.

[27] Ron Kohavi. Scaling up the accuracy of naive-bayes classifiers: a decision-tree hybrid. Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, pages 202–207, 1996.

[28] Lukasz Kurgan and Petr Musilek. A survey of knowledge discovery and data mining process models. The Knowledge Engineering Review, 21(1):1–24, 2006.

[29] Maria Lamas and Constanca Paúl. O envelhecimento do sistema sensorial: implicacões na funcionalidade e qualidade de vida. Actas de Gerontologia, Congresso, 1(1):1–11, 2013.

[30] Lena L. Lim and Ee-Heok Kua. Living alone, loneliness, and psychological well-being of older persons in singapore, 2011. http://www.hindawi.com/journals/cggr/2011/673181/ (Last accessed: Sempter 1, 2014).

[31] Praveen Macherla. Types of web services – big and restful. http://theopentutorials.com/tutorials/web-services/types-of-web-services-big-and-restful/, (Last accessed: August 25, 2014).

[32] Joana Matos. Depressão no idoso. Master's thesis, Faculdade de Medicina da Universidade do Porto, 2010.

[33] Allison Mccann. Smartphone shrink: 5 apps to help your mental health, 2012. http://www.popularmechanics.com/science/health/med-tech/smartphone-shrink-5-apps-to-help-your-mental-health#slide-1, (Last accessed: April 14,

2014).

[34] Larissa Melo and Aletéia Ferruzzi. Depressão na terceira idade, 2013. http://psicologado.com/psicopatologia/transtornos-psiquicos/depressao-na-terceira-idade, (Last accessed: April 8, 2014).

[35] Tom Mitchell. Machine Learning. McGraw-Hill Science/Engineering/Math, 1997.

[36] Tom Mitchell. The discipline of machine learning. pages 1–7, 2006.

[37] Roman MW and Callen BL. Screening instruments for older adult depressive disorders; updating the evidence-based toolbox. Issues Ment Health Nurs, 29:942–41, 2008.

[38] MyM3. Depression calculator, 2014. https://itunes.apple.com/gb/app/depression-calculator/id517937129?mt=8, (Last accessed: April 16, 2014).

[39] Paul Nussbaum. Handbook of Neuropsychology and Aging, chapter 5. Springer, 1997.

[40] Patient.co.uk. Take control of your menthal health, 2014. http://whatsmym3.com/AboutAssessment.aspx, (Last accessed: April 17, 2014).

[41] Fraunhofer Portugal. Smart companion. http://www.fraunhofer.pt/en/fraunhofer_aicos/projects/internal_research/smart_companion.html, (Last acessed: August 8, 2014).

[42] Paul W. Power. The psychological and social impact of illness and disability., 2007. https://books.google.pt/books?id=_KQYiOa9GDoC&dq=rowe+and+kahn+health+promotion+1998&hl=pt-PT&source=gbs_navlinks_s.

[43] Dave Raggett. Raggett on html 4. http://www.w3.org/People/Raggett/book4/ch02.html, (Last accessed: July 25, 2014).

[44] Trygve M. H. Reenskaug. Mvc. http://heim.ifi.uio.no/~trygver/themes/mvc/mvc-index.html, (Last accessed: August 7, 2014).

[45] Nino Rocha and Ana Vasconcelos. Depsigns - depression signs detection through smartphone usage data analysis. Broader, Bigger, Better – AAL solutions for Europe - Proceedings of the AAL Forum 2014 Bucharest, pages 1–6, 2014.

[46] rt.com. Any hang-ups? new phone app logs mood from voice, 2013. http://rt.com/news/smartphone-app-mood-voice-281/, (Last accessed: April 14, 2014).

[47] Johannes Seiler. App may signal cellphone dependency, 2014. http://www3.uni-bonn.de/Press-releases/app-may-signal-cellphone-dependency, (Last accessed: April 15, 2014).

[48] Phill Simon. Too Big to Ignore : The Business Case for Big Data. John Wiley Sons, 2013.

[49] Stanislav Sýkora. Generalized means and averages. www.ebyte.it/library/docs/math09/Means_Heronian.html, (Last accessed: June 10, 2014).

[50] Zaldy S. Tan. Age-proof your mind - detect, delay, and prevent memory loss–before it's too late., 2007. https://books.google.pt/books?id=RVc2AQAAQBAJ&dq=successful+aging+macarthur+foundation&hl=pt-PT&source=gbs_navlinks_s, (Last accessed: September 3, 2014).

[51] Liliana Teixeira. Solidão,depressão e qualidade de vida em idosos: Um estudo avaliativo exploratório e implementação - piloto de um programa de intervenção. Master's thesis, Faculdade de Psicologia da Universidade de Lisboa, 2010.

[52] Roman Timofeev. Classification and regression tress, (cart), theory and application. http://edoc.hu-berlin.de/master/timofeev-roman-2004-12-20/PDF/timofeev.pdf, (Last accessed: August 3, 2014).

[53] Chris Ullman and Lucinda Dykes. Beginning ajax. http://www.wrox.com/WileyCDA/Section/id-303217.html, (Last accessed: July 28, 2014).

[54] Hugh Watson. The crisp-dm model: The new blueprint for data mining. Journal of Data Warehousing, 5(4):13–22, 2000.

[55] WebMD. Alcohol and depression, 2014. http://www.webmd.com/depression/alcohol-and-depresssion?page=2, (Last accessed: May 12, 2014).